A System towards Indian Postal Automation

K. Roy
C.V.P.R Unit, I.S.I,
Kolkata-108; India,
Kaushik_mrg@hotmail.com

S. Vajda
LORIA Research
Center, B.P. 239
54506, Nancy, France
Szilard.Vajda@loria.fr

U. Pal
C.V.P.R Unit, I.S.I,
Kolkata-108; India,
umapada@isical.ac.in

B. B. Chaudhuri
C.V.P.R Unit, I.S.I,
Kolkata-108; India,
bbc@isical.ac.in

Abstract

In this paper, we present a system towards Indian postal automation. In the proposed system, at first, using Run Length Smoothing Algorithm (RLSA), we decompose the image into blocks. Based on the black pixel density and number of components inside a block, non-text block (postal stamp, postal seal etc.) are detected. Using positional information, the destination address block (DAB) is identified from text block. Next, pin-code box from the DAB is detected and numerals from the pin-code box are extracted. Since India is a multi-lingual and multi-script country, the address part may be written by combination of two languages: Arabic and a local language. For the sorting of postal documents written in Arabic and a local language Bangla, a two-stage MLP based classifier is employed to recognise Bangla and Arabic numerals. At present, the accuracy of the handwritten numeral recognition module is 92.10%.

1. Introduction

Postal automation is a topic of research interest for last two decades and many pieces of published article are available towards postal automation of non-Indian languages documents [1-6]. Several systems are also available for postal automation in USA, UK, France, Canada and Australia. But no work has been done towards the automation of Indian postal system.

One of the important tasks in postal automation is to locate destination address block (DAB) and to extract the pin-code from the address part. In India, postal codes (pin-codes) are six digit numbers uniquely specifying a postal zone. There are several difficulties in locating DAB on the envelope because an envelope is composed of not only DAB but also several other meaningful blocks such as return address block, postage stamp block, graphics etc. Furthermore, there exist wide variation due to several kinds of writing instruments, writing habits, the document surface feature and format of the different postal documents. Detection of pin-code from the DAB is also a difficult problem. In some Indian postal documents there are pin-code boxes e.g. post-card, inland letters etc. Also, there exist Indian postal documents without printed pin-code box e.g. ordinary envelope, business letter etc. From the study it is noted that some people write pin-code outside the pre-printed pin-code box area of post-card, inland-letters etc.

System development towards postal automation for a country like India is more difficult than such problem of other country because of its multi-lingual and multi-script behaviour. An Indian postal document may be written by any of the 18 official languages of India. Moreover, some people write the destination address part of a letter in two or more language scripts. For example, see Fig. 2(a), where the destination address is written partly in Bangla and Arabic. Here, the pin-code is written in Arabic while the rest of the address part is written in Bangla. Thus, development of Indian postal automation system is a challenging problem. In this paper, we propose a system towards Indian postal automation where at first, using run length smoothing approach and characteristics of different component, the postal stamp/seal parts are detected and removed from the documents. Next, based on the positional information DAB region is located. Then pin-code from the pin-code box is extracted. Finally, based on two-stage neural network the Bangla and Arabic numerals of the pin-code part are recognized. Bangla is the second most popular language in India and fifth most popular language in the world. Examples of Bangla numerals are shown in Fig. 1 to get an idea of handwriting variability of Bangla numerals.

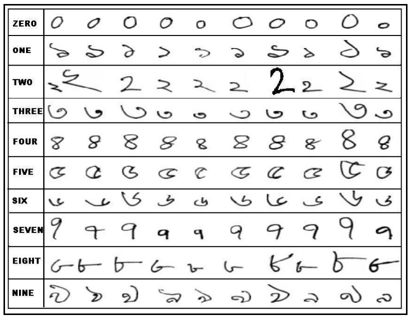


Fig. 1. Sample of Bangla handwritten numerals.

Rest of the paper is organized as follows. Preprocessing including data collection, noise removal, postal stamp detection and deletion, DAB location, pin-code box detection and pin-code extraction are described in Section 2. Section 3 deals with the recognition techniques of the pin-code numerals. Finally, experimental results are provided in Section 4.

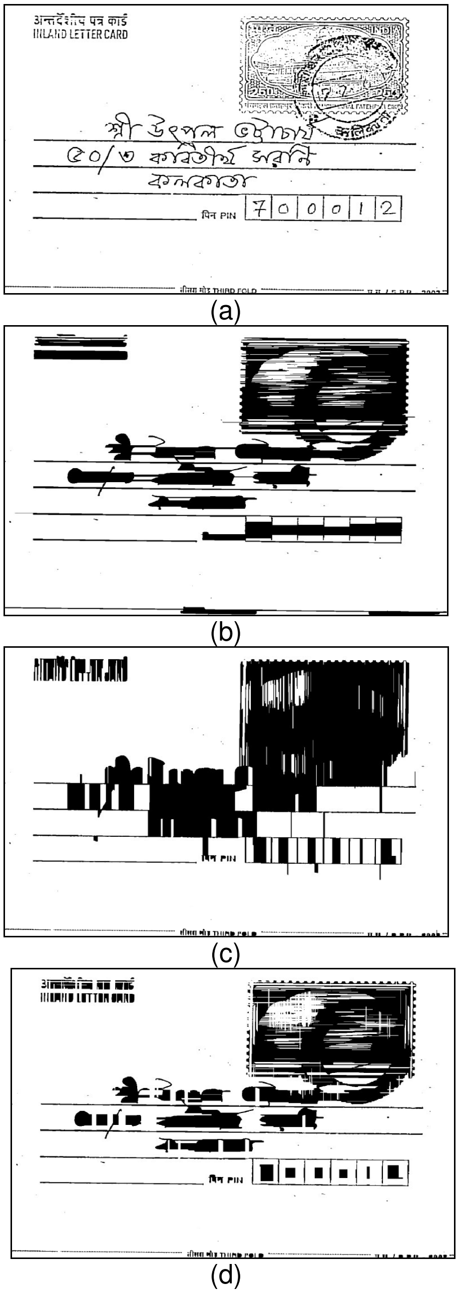
2. Preprocessing

2.1. Data collection and noise removal

Document digitization for the present work has been done from real life data collected from a post-office (Cossipore post office of North Kolkata circle, West Bengal, India). We used a flatbed scanner (manufactured by UMAX, Model AstraSlim) for digitization. The images are in gray tone and digitized at 300 dpi and stored as Tagged Information File (TIF) Format. We have used a two-stage approach to convert them into two-tone (0 and 1) images. In the first stage a pre-binarization is done using a local window based algorithm in order to get an idea of different regions of interest [7]. On the pre-binarized image, Run Length Smoothing Algorithm (RLSA) is applied to overcome the limitations of the local binarized method used earlier. There are more powerful algorithm than RLSA, e.g. Nishiwaki et al.[8] that take care of noisy document. But at present we assume our documents are fairly clean. After this, using component labelling, we select each component and map them in the original image and the final binarized image is obtained using a histogram based global binarizing algorithm on the components [9] (Here '1' represents object pixel and '0' represents background pixel). The digitized document images may be skewed and we used Hough transform to de-skew the documents. The digitized image may contain spurious noise pixels and irregularities on the boundary of the characters, leading to undesired effects on the system. Also, to improve recognition performance, broken numerals should be connected. For pre-processing we use the method due to Chaudhuri and Pal [9].

2.2. Postal stamp detection and deletion

The binary image is processed to extract the Postal stamps and other graphics part present in the image. There are many techniques for text/graphics separation. Here we used a combined technique for the purpose. At first, simple horizontal and vertical smoothing operations of RLSA are performed [10]. The two smoothing results are then combined in a logical AND operation. The results after horizontal, vertical and logical AND operation of Fig. 2(a) are shown in Fig.2 (b), (c) and (d), respectively. The result of logical AND operation is further smoothed to delete the *stray* part (see Fig. 3(a)). On this smoothed image we apply component analysis to get individual blocks. Each smoothed block is then checked for postal stamp/seal block. For each block component we find the boundary of the component and check the density of black pixels over the corresponding boundary area on the original image. We note that for postal stamp/seal block



(a)

(b)

(c)

(d)

Fig. 2: (a) An example of postal document image obtained from an Inland letter. (b) Horizontal run-length smoothing of fig. 2(a). (c) Vertical run-length smoothing of fig. 2(a). (d) Logical AND of 2(b) and 2(c).

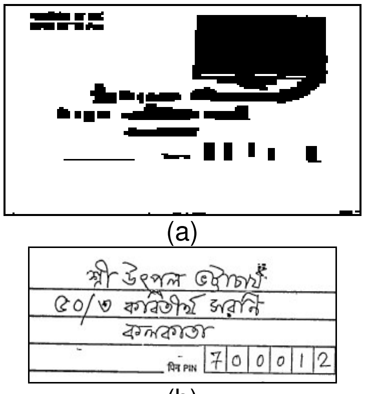
the density of black pixels are very high compared to text line block. Also, we noticed that the postal stamp/seal block contains many small components whereas such small components are not present in other blocks. Based on the above criteria non-text parts are detected. After detection of a postal stamp/seal block we delete that block from the documents for future processing.

2.3. DAB detection

Using positional information of the text block we detect DAB from a postal image. In case of Indian postal documents, address on the postal document is generally written in the manner that DAB will be in the right lower part of the documents. Using this clue we segment DAB from the postal documents.

2.4. Pin-code box detection and extraction

In some Indian postal documents (e.g. Post-card, Inland letters etc.) there are pre-printed boxes to write pin-code. We call these boxes as pin-code boxes. People generally write the destination pin-code inside these boxes. Here, at first, we detect whether there is a pin-code box or not. If it exists, our method will extract the pin-code from the box if pin-code is written within the pin-code box.



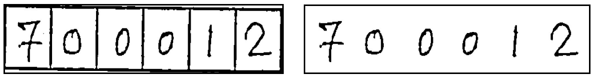
(a)

(b)

Fig. 3: (a) Smoothed version of Fig. 2(d). (b) Detected DAB part.

For pin-code box extraction we apply component labelling and select those components as candidates, which satisfy the following criteria. A component is selected as candidate component if the length of the component is greater than five times the width of the component and the length of the component is less than seven times the width of the component. Since an Indian pin-code box contains six square boxes, the length of a box component will be about six times the width of the component. Based on this principle we choose the candidate component. Let X be the set of these selected components. If we get only one such component then that is considered as the pin-code box. If no such candidate component is obtained, then we assume that there is no pin-code box. If the number of candidate components is two or more, then we decide the best component for pin-code as follows. We scan each column of a selected component from top and as soon as we get a black pixel we stop and note the row value of this point. Let t_i be the row value of the ith column obtained during top scanning. Similarly, we scan each column of the selected component from bottom and as soon as we get a black pixel we stop and note the row value of this point. Let b_i be the row value of the ith column obtained during scanning from bottom. We compute the absolute value of $b_i – t_i$, for all columns. Let W be the width of the component. The selected component satisfying $|(b_i – t_i)– 2R_w| \leq W \leq |(b_i – t_i)+ 2R_w|$ is chosen as pin-code box component. Here R_w is the length of most frequently occurring black run of a component. In other words, R_w is the statistical mode of the black run lengths of the components. The value of R_w is calculated as follows. The component is scanned both horizontally and vertically. Let from this component we get n different run-lengths r_1, r_2,..r_n with frequencies f_1, f_2 ...f_n, respectively. In this case, the value of $R_w = r_i$ where $f_i = \max (f_j)$, j = 1...n.

After detection of the pin-code box, vertical and horizontal lines are detected and deleted. Next depending on the positions of the vertical lines the pin-code numerals are extracted from left to right to preserve the order of occurrence of the numerals. Pin-code box extracted from Fig. 3(b) by the proposed algorithm is shown in Fig. 4(a). Also pin-code numerals extracted from Fig. 4(a) are shown in Fig. 4(b). From experiment of 4200 data we noticed that about 9.5% of the numerals touched/crossed the border of the pin-code box. However, our method can extract most of such cases properly.



(a) (b)

Fig. 4: (a) Extracted part of pin-code box from the DAB shown in Fig. 3(b). (b) Extracted pin-code numerals from the pin-code box.

3. Numeral recognition

After extraction of numerals from the image we proceed for their recognition. For recognition we do not compute any feature from the image. Only the raw images are used for classification. As we have used Neural Network for recognition, which uses a fixed set of input, we normalized the image first to a 28x28 pixel size.

3.1. Normalization

Normalization is one of the important pre-processing factors for character recognition. Normally, in normalization the character image is linearly mapped onto a standard plane by interpolation/extrapolation. The size and position of character is controlled such that the length and width of normalized plane are filled. By linear mapping, the character shape is not only deformed but also the aspect ratio changes. Here we use an Aspect Ratio Adaptive Normalization (ARAN) technique for the purpose [11].

3.1.1. Implementation of normalization. For ease of classification, the length and width of a normalized image plane is fixed. In ARAN adopted by us, however, the image plane are not necessarily filled. Depending on the aspect ratio, the normalized image is centered in the plane with one dimension filled. Assume the standard plane is square and the side length is denoted by L. If the width and height of the input image are W_1 and H_1 respectively then the aspect ratio (R_1) is defined by

$$R_1 = \begin{cases} W_1/H_1 & \text{If } W_1<H_1 \\ H_1/W_1 & \text{otherwise} \end{cases} \qquad (1)$$

3.1.2. Aspect ratio mapping. To implement the normalization, the width and height of the normalized image, W_2 and H_2, are determined. We set max (W_2, H_2) equal to the side length L of the standard plane, while min (W_2, H_2) is determined by its aspect ratio. The aspect ratio of the normalized image is adaptable to that of the original image. Hence the aspect ratio mapping function determines the size and shape of the normalized image. The image plane is expanded or trimmed so as to fit this range. The aspect ratio of the original image is calculated by Eq. (1). To calculate the mapping function (R_2) for the normalized image, we have used square root of the aspect ratio of the original image, given by

$$R2=\sqrt{R_1}.$$

To map the image f (x, y) to the new image g ($x^/$, $y^/$) we have used forward mapping to implement the normalization given by

$$x^/=\alpha x \qquad y^/=\beta y$$

where $\alpha= W_2 /W_1$ and $\beta= R_{2*} H_2/H_1$ if ($W_2> H_2$).
and $\alpha= R_{2*} W_2/W_1$ and $\beta= H_2 /H_1$ otherwise

An example of original and normalized image is shown in Fig. 5(a-b).

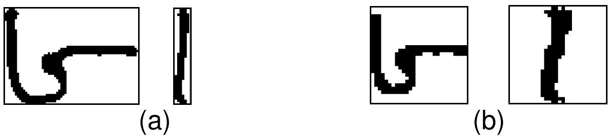

(a) (b)

Fig. 5: (a) Original Image (b) Normalized Image

3.2. Neural network

Based on the above normalization we use Multilayer Perceptron (MLP) Neural Network based scheme for the recognition of Arabic and Bangla numerals [12]. The MLP is, in general, a layered feed-forward network, that can be represented by a directed acyclic graph. Each node in the graph stands for an artificial neuron of the MLP, and the labels in each directed arc denote the strength of synaptic connection between two neurons and the direction of the signal flow in the MLP.

For pattern classification, the number of neurons in the input layer of an MLP is determined by the number of features selected for representing the relevant patterns in the feature space and output layer by the number of classes in which the input data belongs. The Neurons in hidden and output layers compute the sigmoidal function on the sum of the products of input values and weight values of the corresponding connections to each neuron.

Training process of an MLP involves tuning the strengths of its synaptic connections so that it can respond appropriately to every input taken from the training set. The number of hidden layers and the number of neurons in a hidden layer required to design an MLP are also determined during its training. Training process incorporates learning ability in an MLP. Generalization ability of an MLP is tested by checking its responses to input patterns which do not belong to the training set.

Back propagation algorithm, which uses patterns of known classes to constitute the training set, represents a *supervised learning* method. After supplying each training pattern to the MLP, it computes the sum of the squared errors at the output layer and adjusts the weight values of the synaptic connections to minimize the error sum. Weight values are adjusted by propagating the error sum from the output layer to the input layer.

The present work selects a 2-layer perceptron for the handwritten digit recognition. The number of neurons in input and output layers of the perceptron is set to 784 and 16, respectively. This is because the size of the normalized image is 28x28 (784), and the number of possible classes in handwritten numerals for the present case is 16. Although because of bi-lingual (Arabic and local language Bangla) nature of the Indian postal documents the number of numeral class is supposed to be 20, we have used only 16-classes in the output layer of the MLP. This is because Arabic and Bangla 'zero' are (historically the Arabs borrowed the zero from India and transported to the west) same and we consider these two as a single class. Also Arabic 'eight' and Bangla 'four' are same. Arabic and Bangla 'two' are very similar. Arabic 'nine' and Bangla 'seven' are also similar. To get an idea of such similarity see Fig. 6.

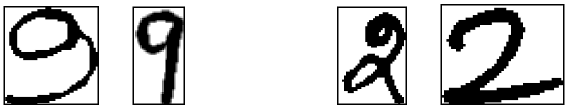

(a) (b)

Fig. 6: (a) Arabic Nine and Bangla Seven, (b) Arabic and Bangla Two.

The number of hidden units of the proposed network is 400, Back Propagation learning rate is set to suitable values based on trial runs. The *stopping criteria* of BP algorithm selected for the present work is that the sum of the squared errors for all of the training patterns will be less than a certain limit.

In the proposed system we used three classifiers for the recognition. The first classifier deals with 16-class problem for simultaneous recognition of Bangla and Arabic numerals. Other two classifiers are for recognition of Bangla and Arabic numerals, separately. The Bangla classifier is developed only for 10 Bangla numerals and the Arabic classifier is developed for 10 Arabic numerals. Based on the output of the 16-class classifier we decide the language in which pin-code is written. As mentioned earlier, Indian pin-code contains six digits. If out of these six numerals majority of the numerals are recognised as Bangla by the 16-class classifier then we use Bangla classifier on this pin-code to get higher recognition rate. Similarly, if the majority of the numerals are recognised as Arabic by the 16-class classifier then we use Arabic classifier on this pin-code to get better result.

4. Result and discussion

4.1. Result on DAB detection

The performance of the proposed system on postal stamp/seal detection, and DAB location are as follows. We have tested our system on 2860 postal images and noted that the accuracy for postal stamp/seal detection, and DAB location are 95.98% and 98.55%, respectively. Some errors in postal stamp/seal detection and DAB location appeared due to overlapping of postal stamp/seal with the text portion of address part. Some errors also appeared due to poor quality of the images.

4.2. Result on pin-code box detection

The performance of the proposed system on pin-code box extraction is as follows. We have tested our system on 2860 postal images and the accuracy for pin-code box extraction module is 97.64%. The main source of errors was due to broken pin-code box, poor quality of the images and touching of the text portion of DAB with the pin-code box.

4.3. Result on numeral recognition

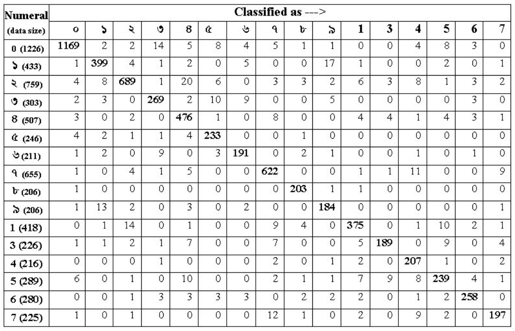
For the experiment of the proposed numeral recognition approach we collected 7500 postal documents images. Some form documents were also considered for data collection. We collected 15096 numerals from these documents for experiment. 80% of the data were collected from postal documents and the rest were from form documents. Among these numerals 8690 (4690 of Bangla and 4000 of Arabic) were selected for training of the proposed 16-class recognition system and the remaining 6406 (3179 of Bangla and 3227 of Arabic) numerals were used as test set. For experiment on Arabic and Bangla individual classifier we also collected two datasets of 10677 and 11042 numerals. We consider 5876 (6290) data for training and 4801 (4752) data for testing of Arabic (Bangla) classifiers.

The overall accuracy of the proposed 16-class classifier and individual Bangla and Arabic classifiers on the above data set are given in Table 1. From the Table we note that in Bangla classifier we obtained 2.03% better accuracy than the 16-class classifier. This is due to decrease in the number of classes and also decrease in the shape similarity among Arabic and Bangla numerals. The, confusion matrix of three classifiers are shown in Table 2, Table 3(a) and 3(b), respectively. In the table the data size of different numerals are not equal. This is because we have collected majority of the data from postal documents and the numerals in postal documents are not equally distributed.

Table 1: Overall numeral recognition accuracy on the training and test set of data.

Classifier	Recognition rate	
	Training Set	Test Set
16-class classifier	98.31%	92.10%
Bangla classifier	98.71%	94.13%
Arabic classifier	98.50%	93.00%

Table 2: confusion matrix obtained for 16-class classifier.



Numeral	Classified as --->															
(data size)	০	১	২	৩	৪	৫	৬	৭	৮	৯	1	3	4	5	6	7
0 (1226)	1169	2	2	14	5	8	4	5	1	1	0	0	4	8	3	0
১ (433)	1	399	4	1	2	0	5	0	0	17	1	0	0	2	0	1
২ (759)	4	8	689	1	20	6	0	3	3	2	6	3	8	1	3	2
৩ (303)	2	3	0	269	2	10	9	0	0	5	0	0	0	0	3	0
৪ (507)	3	0	2	0	476	1	0	8	0	0	4	4	1	4	3	1
৫ (246)	4	2	1	1	4	233	0	0	1	0	0	0	0	0	0	0
৬ (211)	1	2	0	9	0	3	191	0	2	1	0	0	1	0	1	0
৭ (655)	1	0	4	1	5	0	0	622	0	0	1	1	11	0	0	9
৮ (206)	1	0	0	0	0	0	0	0	203	1	1	0	0	0	0	0
৯ (206)	1	13	2	0	3	0	2	0	0	184	0	0	0	0	0	1
1 (418)	0	1	14	0	1	0	0	9	4	0	375	0	1	10	2	1
3 (226)	1	1	2	1	7	0	0	7	0	0	5	189	0	9	0	4
4 (216)	0	0	0	1	0	0	0	2	0	1	2	0	207	1	0	2
5 (289)	6	0	1	0	10	0	0	2	1	1	7	9	8	239	4	1
6 (280)	0	0	1	3	3	3	3	0	2	2	2	0	1	2	258	0
7 (225)	1	0	1	0	0	0	0	12	1	0	2	0	9	2	0	197

From table 2 (Table 3(a)) it can be noted that highest accuracy is obtained on Bangla numeral ‘eight’ 98.54% (98.06%). For Arabic numerals classifier we noted that highest accuracy is obtained for numeral ‘zero’ (97.63%). Although the results of the Arabic classifier on Indian pin-code is only 93.0%, which is not attractive, we test this system on the MNIST data set to get a comparative result. From the experiment, we noticed that from MNIST database we obtained 98.5% accuracy on Arabic classifier. Low accuracy on Indian postal documents is due to variability of handwritings as well as poor postal documents and bad writing medium.

From the experiment we noted that the most confusing numeral pair was Bangla 'one' and Bangla 'nine' (shown in Fig. 7(a)); they confuse in about 6.3% cases. Their similar shapes rank the confusion rate at the top position. Second confusion pair is Bangla seven and Arabic seven (see Fig. 7 (b)) with confusing rate 5.3%.

We did not incorporate any rejection scheme in the proposed system, which we plan to add in future.

Table 3: Confusion matrix obtained (a) Bangla and (b) Arabic classifier.

(a)

Numeral	Classified as --->									
(data size)	০	১	২	৩	৪	৫	৬	৭	৮	৯
0 (1226)	**1175**	3	1	14	4	10	1	10	0	8
১ (433)	2	**394**	4	1	2	0	6	0	0	24
২ (759)	5	6	**705**	2	20	7	1	3	6	4
৩ (303)	3	4	0	**270**	1	10	10	0	0	5
৪ (507)	5	1	6	0	**480**	4	0	10	0	1
৫ (246)	5	2	2	0	1	**232**	2	0	1	1
৬ (211)	1	2	1	11	0	3	**189**	0	2	2
৭ (655)	1	0	6	1	5	2	0	**640**	0	0
৮ (206)	1	0	0	0	0	0	0	0	**202**	3
৯ (206)	2	11	1	0	3	0	3	0	0	**186**

(b)

Numeral	Classified as --->									
(data size)	0	1	2	3	4	5	6	7	8	9
0 (1226)	**1197**	0	0	2	7	7	3	0	3	7
1 (418)	0	**369**	16	1	3	12	3	4	0	10
2 (759)	7	6	**701**	7	8	0	6	2	17	5
3 (226)	4	5	4	**192**	1	4	0	5	6	5
4 (216)	0	1	0	0	**208**	1	1	3	0	2
5 (289)	7	8	3	14	6	**234**	4	1	11	1
6 (280)	2	1	4	0	1	1	**268**	0	3	0
7 (225)	0	3	1	0	9	1	0	**200**	0	11
8 (507)	4	4	5	4	2	4	3	1	**475**	5
9 (655)	2	3	4	1	14	0	1	7	2	**621**

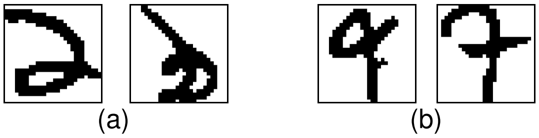


Fig. 7: Examples of some confused handwritten numeral pairs. (a) Bangla one and nine (b) Bangla seven and Arabic seven.

5. Conclusion

A system towards Indian postal automation is discussed here. In the proposed system, at first, using RLSA, we decompose the image into blocks. Based on the black pixel density and number of components inside a block, non-text block (postal stamp, postal seal etc.) are detected. Using positional information, the DAB is identified from text block. Next, pin-code box from the DAB is detected and numerals from the pin-code box are extracted. Finally pin-code digits are recognised for postal sorting according to the pin-code of the documents. This is the first report of its kind and hence we cannot compare the results of different modules of the proposed system.

Acknowledgement: Partial financial support by Indo-French Centre for the Promotion of Advanced Research (IFCPAR) is greatly acknowledged.

6. References

[1] R. Plamondon and S. N. Srihari, “On-line and off-line handwritten recognition: A comprehensive survey”, IEEE Trans. on PAMI, Vol. 22, pp. 62-84, 2000.

[2] U. Mahadevan, and S. N. Srihari, “Parsing and Recognition of City, State, and ZIP Codes in Handwritten Addresses”, In Proc. of Fifth ICDAR, pp. 325-328, 1999.

[3] X. Wang, and T. Tsutsumida, “A New Method of Character Line Extraction from Mixed-unformatted Document Image for Japanese Mail Address Recognition”, In Proc. of Fifth ICDAR, pp. 769-772, 1999.

[4] D. Bartnik, V. Govindaraju, S. N. Srihari and B. Phan, “Reply Card Mail Processing”, In Proc. of ICPR, pp. 633-636, 1998.

[5] G. Kim, and V. Govindaraju, “Handwritten Phrase Recognition as Applied to Street Name Images”, Pattern Recognition, Vol. 31, pp. 41-51, 1998.

[6] S. N. Srihari, and E.J. Keubert, “Integration of Hand-Written Address Interpretation Technology into the United States Postal Service Remote Computer Reader System”, In Proc. of Forth ICDAR, pp. 892-896. 1997.

[7] P. Palumbo, P. Swaminathan, and S. Palumbo, “Document Image Binarization: Evaluation of Algorithms”, SPIE Applications of Digital Image Processing IX, Vol. 697, pp. 278-285, 1986.

[8] D. Nishiwaki et al, “Robust Frame Extraction and Removal for Processing Form Documents”, Graphics Recognition Algorithms and Applications, LNCS Vol. 2390, pp. 46-66, 2002.

[9] B. B. Chaudhuri and U. Pal, “A complete printed Bangla OCR system”, Pattern Recognition, Vol. 31, pp. 531-549, 1998.

[10] F. M. Wahl, K. Y. Wong, R. G. Casey, “Block segmentation and text extraction in mixed text / image documents", Computer Graphics and Image Processing, Vol. 20, pp. 375 - 390, 1982.

[11] Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, Hiromichi Fijisawa, “Handwritten digit recognition: investigation of normalization and feature extraction techniques”, Pattern Recognition, Vol. 37, pp. 265-278, 2004.

[12] E. William Weidemen, T. Michael Manry, Hung-Chun Yau and Wei Gong, “Comparisons of a neural network and a nearest-neighbour classifier via the numeric handprint recognition problem”, IEEE Trans. on Neural Network, Vol. 6, pp. 1524-1530,1995.