An Automated Mail Sorter System using SVM Classifier

Arun K.S.
Assistant Professor
Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, India

Jerin Thomas
Assistant Professor
Computer Science & Engineering
Amal Jyothi College of Engineering
Kanjirappally, India

ABSTRACT

This paper proposes an Automated Mail Sorter (AMS) system that scans a mail and interprets one of the imperative fields of the destination address, the pin code to sort the mails. The scanned document was segmented into different fields to extract the pin code. A general classifier for the recognition of pin-code digits written in English was then employed. The recognition system consists of a feature extractor and a classification network. The feature extracted was Hu moments and the classification network was the support vector machine using a polynomial kernel of order 2. Other classification networks such as the multi feature recognizer and decomposing network were also used, but SVM gave the maximum accuracy.

General Terms

Pattern Recognition, Image Segmentation

Keywords

Support Vector Machine, Hu Moments, RLSA Algorithm, Connected Component Labeling, Region Labeling.

1. INTRODUCTION

Because of its multi-lingual and multi-script behavior, postal automation system development for a country like India is more difficult than such a task in other countries. Based on the six digit pin code number any post office in India can be located. Because of educational backgrounds there is a wide variation in writing style and medium, hence the development of Indian postal address reading system is a challenging problem.

Hong Yan et al developed a high pass filter approach for locating address blocks and postcodes in mail-piece images [1]. In this approach an image was segmented into clusters and the candidate cluster based on size, position and pixel density consistency is selected for subsequent processing by OCR methods. Tatsuhiko Kagehiro et al. proposed an address-block extraction technique using Bayesian rule for Japanese mail [2]. Their method was composed of two steps, nomination of address block candidates and evaluation of these candidates by using the Bayesian rule according to each of address-block type. There are various types of address-block in Japanese mail hence a single set of parameters for the candidate cannot be determined. So several address-block dictionaries are prepared according to each type and an address-block candidate for each of type of address is assumed. Finally a confidence value for each type is calculated independently as each dictionary is selected and referenced.

Yonekura et al proposed a postal envelope segmentation method based on 2-D histogram clustering and watershed transform segmentation task consists in detecting the modes associated with homogeneous regions in envelope images such as handwritten address block, postmarks, stamps and background [3]. The homogeneous modes in 2-D histogram are segmented through the morphological watershed transform. The advantage of this method is that very little a priori knowledge of the envelope images is required. All the above pieces of work were done for script separation from the scanned documents. Similarly works on the identification of hand-written script were also existing. Kimura et al proposed a novel method for ZIP code recognition based on word recognition algorithm, where a numeral string is recognized as a word [4]. Filatov et al proposed a graph based handwritten digit string recognition based on matching input sub graphs with prototype symbol graph [5]. The search for a match between the input sub graph and prototype graph is conducted using a set of transformation.

.

2. METHODOLOGY

The proposed scheme of postal automation consists of two modules, namely pin code extraction and digit recognition. Pin code extraction module is an integration of five stages namely, preprocessing, segmentation, region labeling, address block extraction and line and word extraction. For digit recognition affine invariant Hu moments were obtained from the extracted pin code and given to an SVM with polynomial kernel of order 2 for the classification.

At first, non-text blocks like postal stamp, postal seal etc are detected. Then destination address block (DAB) is identified from the postal document and followed by lines and words of the DAB are segmented. The postal documents in India contain two languages. English and regional language. This paper aims at automation of English inlands only. Indian postal code is a six-digit number. Based on this six-digit pin-code we can locate a particular post office in a village.

In preprocessing stage the given postal document is binarized using Otsu’s method. Connected component analysis and run length smearing algorithm (RLSA) were used to segment the binarized postal document in segmentation stage [5]. The region labeling gives a unique label to each segment in the given document. Positional information was used to extract the address block. Finally horizontal and vertical projection profiles were used to extract lines and words respectively.

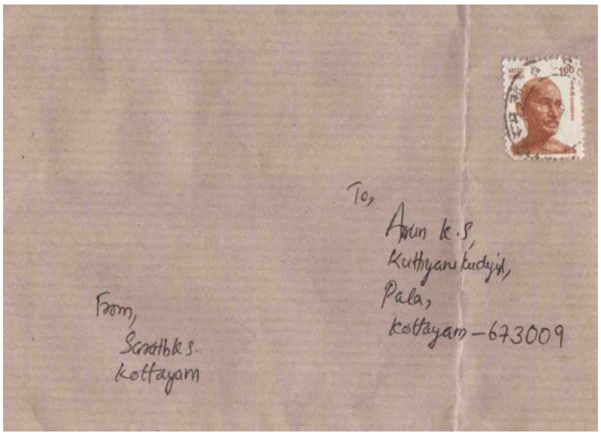

Fig 1: Sample Input Data

3. PREPROCESSING

Document digitization was done using a flatbed scanner. The images were in gray tone with 300 dpi and stored as JPEG Format. The preprocessing step involves the binarization of the image. Binarization is done by Otsus method, the algorithm of which is detailed below [4]:

1. The document is scanned for the maximum intensity pixel and minimum intensity pixel.

2. The mean value is found.

3. It is checked weather the number of grey levels below and above threshold is same.

4. If not mean value is found by taking the mean of above and below values and finding the average.

5. This process is repeated until the difference between previous and new threshold is small.

4. IMAGE SEGMENTATION

Image segmentation is a useful operation in many image processing applications. Connected Component Analysis extract regions which are not separated by a boundary. Any set of pixels which is not separated by a boundary is call connected. Each maximal region of connected pixels is called a connected component. The set of connected components partition an image into segments.

4.1 Run Length Smearing Algorithm

The run-length smearing algorithm (RLSA) works on binary images where white pixels are represented by 0s and black pixels by 1s [5]. The algorithm transforms a binary sequence x into y according to the following rules:

1. 0s in x are changed to 1s in y if the number of adjacent 0s is less than or equal to a predefined threshold C.

2. 1s in x are unchanged in y.

These steps have the effect of linking together neighboring black areas that are separated by less than C pixels. The RLSA is applied row-wise to the document using a threshold Ch, and column-wise using threshold Cv, yielding two distinct bitmaps. These two bitmaps are combined in a logical AND operation.

Additional horizontal smearing is done using a smaller threshold, Cs, to obtain the final bitmap. Then, connected component analysis is performed on this bitmap, and using threshold C21 and C22 on the mean run of black pixel in a connected component and block height, connected components are classified into text and non-text zones. The algorithm is fast but presents some limits: the threshold values have to be set a priori, it can be applied only to documents with a rectangular structured layout. In order to get rectangular blocks a post processing step is required. Algorithm for RLSA is detailed below:

1. Scan the document from the top row.
2. Count the number of continuous off pixels in a row.
3. If the number lesser than a threshold (4) convert the pixels to on pixels.
4. Reset the counter and repeat the procedure for all rows.
5. Do scanning column wise.
6. Repeat the same procedure.
7. AND the two results.

The output after applying connected component analysis does not provide complete segmentation of the image. So run length smearing algorithm is performed on the image obtained after connected component analysis. This gives the desired segmentation result as shown in figure 2.

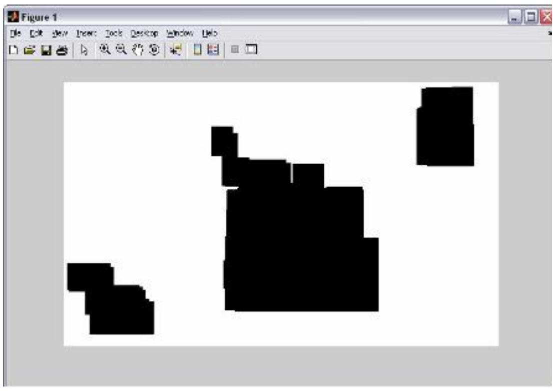

Fig 2: Segmented Image

4.2 Region Labeling

The segmented image after RLSA is labeled for giving a unique number for each segment in the image. Pixels in the same region are given same number[6]. For finding all regions that are connected, the region labeling algorithm is as follows:

1. The binarized image is scanned for a pixel with value 1.

2. Label Number is initialized as 2.

3. The selected pixel is assigned the value Label Number.

4. If any pixel in the eight neighborhood of the pixel with Label Number is 1, then that pixel is assigned value=Label Number.

5. Repeat step 4 until no pixel in the eight neighborhood of pixel labeled Label Number is 1.

6. Check for unlabelled pixel with value 1. If this does not exist, stop. Else repeat steps 4-6.

4.3 Extraction of Address Block

From the labeled postal document the address block is segmented using the following criteria:

1. Regions with width greater than a threshold is selected.

2. The height of such regions should be above threshold.

3. The position of such documents must be towards center.

Regions with higher width are selected based on the threshold value. This includes the DAB, sender's address and region for stamp or any other writings in the inland. The height is then compared with the threshold. Stray writings and stamp section are deleted in this section. The destination address and stamp(if big) are removed by the positional analysis mentioned in step 3. The DAB is the section towards center in both x and y directions

4.4 Extraction of Lines and Words

4.4.1 Projection Profile

Line segmentation is an operation that belongs to the page segmentation class of problems. Although grey level images can be used, in general the segmentation is performed on binarised version of the images. In the current work projection profiles are used for segmentation of lines, words and characters. Lines are extracted by thresholding the projection profile. Here the threshold value is set as 0. This is done on the assumption that no two lines intersect. If the threshold value is slightly increased to 3 or 4, intersecting line segments can also be segmented. Projection profile almost intuitively gives to us the lines of text that are formed by the image. If we imagine a simple vertical line to cross the entire image at the upper side of the projection profile, walk through that hypothetical line from the top until the bottom, we will cross several times, two distinct areas. One area will be empty in the histogram, and the other area will be full. This approach still leaves some cases unattended. Unexpected shorter lines remain undetected. However, this error can be easily corrected if we reapply the same algorithm to the portions of the image that didn't contribute with any line. Thus shorter lines are detected. A second kind of problem arises when we walk through the projection profile, along the mean value. A single line may be cut in two or more 'lines'. This problem is corrected through a voting procedure. Each line votes with the corresponding height to choose a representative. Majority elects the representative lines and those that differ significantly (50%) from the representative are discarded. An example of projection profile is shown

1. Scan the document row wise or column wise and take the count of on pixels in every row or column.

2. From the obtained values fix a minimum threshold value.

3. Segment the image between the two minimum values.

4.4.2 Word Extraction

This process has some steps that are very similar to the problem of line segmentation. The process starts by computing the projection profile of the image, iterating across vertical lines. With this projection profile, the histogram of the space length is build. Two distinct areas are clearly present. The space between characters corresponds to the smaller bin values. The spaces between words correspond to higher bin values. We assume that two characters are separated by a very small amount of pixels. On the other hand, we assume that two words are separated by a significant amount of pixels. With the projection profile we can construct a histogram of the lengths of white spaces that are in the line.
From the extracted words the pin code digits are segmented out by taking the density of pixels in each segment and also by taking the width of each character. The output is as shown in figures 3.14 and 3.15 Individual digits are segmented by taking projection profile

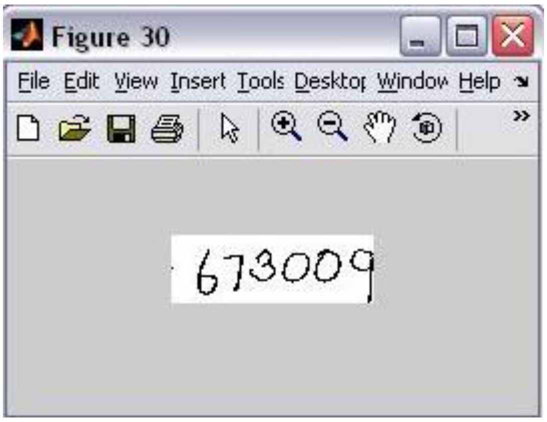

Fig 3: Segmented Pin Code

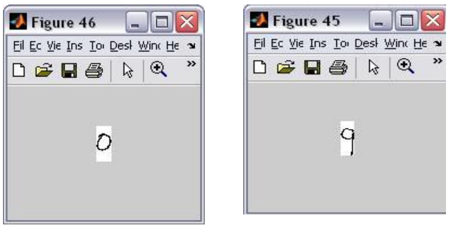
Figure 46
Fil Ec Vie Ins To Desk Winc He
0
Figure 45
Fil Ec Vie Ins To Desk Winc He
9

Fig 4: Segmented Pin Code

5. DIGIT RECOGNITION

5.1 Hu Moments

Invariant moments were introduced by Hu in 1962. Since an object can appear in any position in the image, it is required to shift the image for invariant features. Likewise, since an object must be recognized independently of its angular orientation, we require rotationally invariant features[3]. Finally, for removing magnification in image and other size effects, it is required to scale (or size) the image, when applied to objects which have been segmented into object and background. The Hu moments provide affine transformation invariant moments. The Hu moments of the image are taken as features and given to an SVM with polynomial kernel of order 2 and it classifies.

If image is binary that is one or zero, where 1 corresponds to object and 0 corresponds to background, moments and moment invariants capture the shape of the object silhouette. The image should be free from geometric attacks. The typical geometrical attacks include rotation, scaling and translation of the image. These kinds of attacks can be represented by affine transform. The affine transform with scaling parameters (a,b), rotation angle f and translational parameters (Tx, Ty) can be defined as:

$$\begin{bmatrix} x_a \\ y_a \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix}$$

where (x, y) are the pixel coordinates of an input image and (xa, ya) are the pixel coordinates of a transformed image. The affine transform parameters can be estimated using image moments. For discrete images the two-dimensional moment of order p+q is defined as follows:

$$m_{pq} = \sum_{y=1}^{N_2} \sum_{x=1}^{N_1} x^p y^q I(x,y)$$

In the case of binary Images I(x,y)=0 or 1;

$$m_{pq} = \sum_{y=1}^{N_2} \sum_{x=1}^{N_1} I(x,y)$$

corresponds to area measured in pixels. Considering pixels as point masses, the center of mass of the image (ie.of the segmented output) is given by:

$$\bar{x} = \frac{m_{10}}{m_{00}} \qquad \bar{y} = \frac{m_{01}}{m_{00}}$$

The shift invariance is achieved by calculating the centralized moments by

$$\mu_{pq} = \sum_{y=1}^{N_2} \sum_{x=1}^{N_1} (x-\bar{x})^p (y-\bar{y})^q I(x,y)$$

Scale invariance is achieved by normalizing with respect to area (m00), i.e. if we (notionally) scale the row and column dimensions by:

$$\lambda = \frac{1}{\sqrt{m_{00}}}$$

The normalized moments are:

$$n_{pq} = \frac{m_{pq}}{(\sqrt{m_{00}})^{\gamma}}$$

Finally ,the rotation invariant Hu moments are given by:

$$h_1 = n_{20} + n_{02}$$

$$h_2 = (n_{20} - n_{02})^2 + 4n_{11}{}^2$$

$$h_2 = (n_{20} - n_{02})^2 + 4n_{11}{}^2$$

$$h_3 = (n_{30} + n_{12})^2 + (n_{21} + n_{03})^2$$

5.2 Support Vector Machines

A classification system based on statistical learning theory called the support vector machine has recently been applied to the problem of handwritten digit classification. The general class of algorithms resulting from this process is known as kernel methods or kernel machines [8]. They exploit the mathematical techniques mentioned earlier in order to achieve maximum flexibility, generality and performance, in terms of both generalization and computational cost. They owe their name to one of the central concepts in their design: the notion of kernel functions, used in the representation of the non-linear relations discovered in the data. This technique is said to be independent of the dimensionality of feature space as the main idea behind this classification technique is to separate the classes with a surface that maximize the margin between them, using boundary pixels to create the decision surface. The data points that are closest to the hyper plane are termed support vectors. The Number of support vectors is thus small as they are points close to the class boundaries [9].

One major advantage of support vector classifiers is the use of quadratic programming, which provides global minima only. The absence of local minima is a significant difference from the neural network classifiers. Like neural classifiers, applications of SVM to any classification problem require the determination of several user-defined parameters. Some of these parameters are the choice of a suitable multiclass approach, Choice of an appropriate kernel and related parameters, determination of a suitable value of regularisation parameter (i.e. C) and a suitable optimisation technique. SVMs were initially developed to perform binary classification; though, applications of binary classification are very limited. Most of the practical applications involve multiclass classification, especially in remote sensing land cover classification. A number of methods have been proposed to implement SVMs to produce multiclass classification. Most of the research in generating multiclass support vector classifiers can be divided in two categories. One approach involves in constructing several binary classifiers and combing their results while other approach considers all data in one optimization formulation.

Given a training set of instance label pairs (x_i, y_i), i= 1… l where x_i €R^{un} and y€$(1,-1)^1$, the SVM require the solution of the following optimization problem:

$$Min_{w,b,\varepsilon} \frac{1}{2} w^T w + c \sum_{i=1}^{l} \varepsilon_i$$

$$y_i (w^T \phi(x_i) + b) > 1 - \varepsilon_i,$$

$$\varepsilon_i > 0$$

where C is the capacity constant, w is the vector of coefficients, b a constant. Here training vectors xi are mapped into a higher dimensional space by the function Ø called the Kernel.

5.2.1 SVM for Multiclass Classifier

Originally, SVMs were developed to perform binary classification. However, applications of binary classification are very limited especially in remote sensing land cover classification where most of the classification problems involve more than two classes. A number of methods to generate multiclass SVM from binary SVM's have been proposed by researchers and is still a continuing research topic. The method adopted in this project is one against one approach.

5.2.2 SVM for Multiclass Classifier

In this method, SVM classifiers for all possible pairs of classes are created. Therefore, for M classes, there will be binary classifiers. The output from each classifier in the form of a class label is obtained. The class label that occurs the most is assigned to that point in the data vector. In case of a tie, a tie-breaking strategy may be adopted. A common tie-breaking strategy is to randomly select one of the class labels that are tied. The number of classifiers created by this method is generally much larger than the previous method. However, the number of training data vectors required for each classifier is much smaller. The ratio of training data vector size for one class against another is also. Therefore, this method is considered more symmetric than the one against- the-rest method. Moreover, the memory required to create the kernel matrix is much smaller. For M classes $M(M-1)/2$ classes are required for this approach.

6. RESULTS AND CONCLUSIONS

A system towards Indian postal automation is discussed here. In the proposed system, at first, using RLSA and connected component analysis, the image is decomposed into locks. Region Labeling is performed for labeling the segments. Using positional information, the DAB is identified from the segmented blocks. Then the line and word from the DAB are segmented to extract the pin-code from the DAB and numerals from the pin-code box are extracted. Address block extraction and digit segmentation had high percent of accuracy. Pin-code digits are recognized for postal sorting according to the pin-code of the documents. The digits are recognized using Hu moments as features and SVM as classification network. An accuracy of 99% was obtained for training digits.

7. REFERENCES

[1] K.Roy, S.Vaida,U.Pal,B. B. Chaudhuri and A.Belaid, "A System for Indian Postal Automation", LORIA Research Center,B.P 239 54506, Nancy, France,2002 .

[2] Jameel Ahmed, Essha M. Alkhalifa, "Handwritten Digit Recognition Using an Optimizing Algorithm", Proceedings of the 9th Internaional Conference on Neural Information Processing,2002.

[3] Jonathan Campbell, "Moment Invariant Shape Features: a brief explanation", Letterkenny Institute of Technology, Ireland, 2004.

[4] Decong Yu, Lihong, Ma."Digit Recognition Based on Multi-features", IEEE Systems and Design Engineering Symposium, 2007.

[5] Venu Govindaraju and Sergey Tulyakoy, "Postal Address block location by contour clustering", Proceedings of the Seventh International Conference on Document Analysis and Recognition,2003.

[6] TatsuhikoKagehiro, Masahi Koga, Hiroshi Sako and Hiromichi Fujisawa, Address Block Extraction by Bayesian Rule", Proceedings of the 17th International Conference on Pattern Recognition, 2004.

[7] Adrian P. Whichello and Hong Yan, "Locating Address Blocks and Postcodes in Mail-Piece Images", Proceedings of the 14th International Conference on Pattern Recognition,1996.

[8] J.Weston and C. Watkin,"Multi-class Support Vector Machines", Technical Report, Department of Computer Science, 1998.

[9] Dr Robert Sanderson, "Data Mining", Dept. of Computer Science University of Liverpool, 2008.

[10] Bernard Lemaric, "Practical Implementation of a Radial Basis Function Network for Handwritten Network for Handwritten Digit Recognition", Proceedings of the Second National Conference on Document Analysis and Recognition,1993.