Handwritten Manuscript Digitizer

Kaushil Ruparelia
Student, T. S. E. C.

Ashay Shah
Student, T. S. E. C.

Seema Wadhwani
Student, T. S. E. C.

M. Mani Roja, PhD
Asst. Prof, T. S. E. C.

ABSTRACT

In India, there are various instances where information is gathered by filling a questionnaire or a form. This information is then updated manually into the databases by the concerned authorities. Due to manual data entry, human error results in the capture of inaccurate data and thereby results in faulty storage and analysis of the data. The process is time consuming with a greater probability of error. This document serves as a guideline to automate and expedite the above process.

The paper contains ideas of converting the handwritten samples into electronic data. It uses the kernel method of Multi class Support Vector Machine for handwritten character recognition. The data is first extracted in form of individual images for the corresponding data field, pre processed and converted to digital format. This reduces the time and human effort needed for the same. This paper aims at easing the process of evaluation by automating the correction process.

Keywords

HCR, OCR, Support Vector Machine, Kernel trick.

1. INTRODUCTION

Handwriting recognition is one of the challenging and fascinating researches of image processing and pattern recognition in the recent years. In the current scenario, filling a form or a questionnaire is a mandatory ritual for activities like setting up a bank account, obtaining official documents, medical forms, securing admission in a college, enrolling for memberships and various other similar cases. With human efforts, the details so obtained are fed into a database of an Enterprise Resource Planning (ERP) software.

Procuring details using paper is a time consuming process. It involves extracting the data manually and match it with the corresponding field in the database. Manual efforts are error prone and less reliable due to possibilities of errors in spelling of the data fields. Extracting data from a large number of forms and recording the same in the ERP software numerous times is tiresome and monotonous.

This paper describes preprocessing methods to capture the data to its corresponding field. Offline recognition is performed on a scanned image of handwriting and thus contains no temporal data. The kernel method of Multi class Support Vector machine classifies the handwritten samples and converts them into digital format. The extracted data is stored to its corresponding field in spreadsheet applications like excel.

The primary aim of this paper is at completing the digitization process briskly and in bulk. The described process aims at realizing and extracting prominent information for further analysis. For one use case, after correcting the answer booklets, professors tend to add comments and marks for the section to the left of the sheet. This process can extract and match that data to the respective questions and calculate the total marks and grades thereby making it available for the student on the server.

2. SUPPORT VECTOR MACHINE (SVM)

Machine learning is about learning the structure of the input from the data. In machine learning, support vector machines are supervised learning models [1]. They are associated learning algorithms that analyze data and recognize patterns. In our case it is used to analyze and recognize the patterns of handwritten alphabets and numbers.

2.1 Learning

In order to perform classification, the machine needs to learn from the data sets. This involves collecting sample inputs from the data set and allocating it to its corresponding values. For example, in handwritten character recognition, an image set containing numbers and alphabets is collected from various sources. These letters are then processed to match the corresponding images to its values. The input images are converted to a 2 dimensional matrix. Ideally, the images undergo the process of binarization to improve the performance of the system. This also ensures low memory usage with maximum output. The system is then trained for the samples and its corresponding values. As the training completes, the trained data is stored in a local file system ready to use for Support Vector Machine classification.

3. PROCESS INVOLVED IN DIGITIZATION

The process of procuring data from a questionnaire to a digitized format is carried out by recognition of handwritten characters.

Handwritten character recognition is essentially an application intended at detecting and recognizing characters from an input image and converting it into ASCII or other equivalent machine editable format. The difficulties in the process of digitization are as follows:

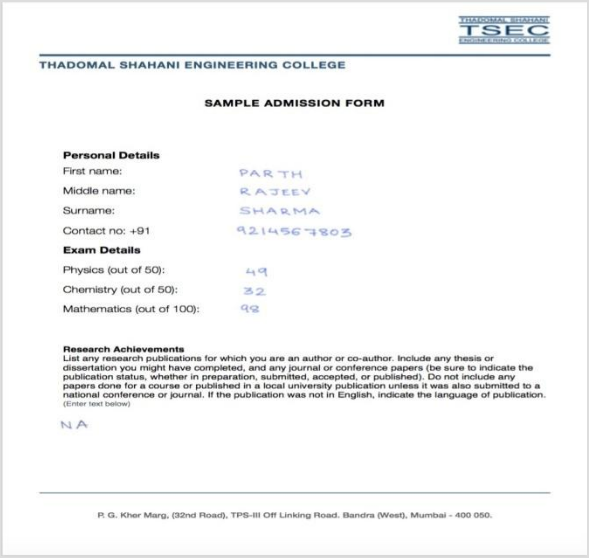
- The ‘same’ character differs in size and shape along with the handwriting style from person to person. The character also differs with time for the same person thus making recognition difficult.
- Like any image, visual characters are subject to spoilage due to noise.

A general statement of the problem of machine recognition of OCR [2] can be formulated as: given a scanned image of a handwritten manuscript, detect and recognize characters from it using a stored database of characters. Available information such as alphabets, digits, special characters may be used in narrowing the search for better accuracy.

The solution to the mentioned problem consists of six stages: Scanning, Image Pre-processing, Segmentation, Feature Extraction, SVM based Classification & Recognition and Analysis & storage.

3.1 Template processing

This is one of the most crucial steps of the process. In this process, the intention is to adjust the image of the questionnaire and align it to the intended template. This ensures that the coordinates stored in the system for processing the data extraction exactly matches with the handwritten data and loss of data is avoided. A processed template of size 954 x 1236 is as shown in figure 1.



THADOMAL SHAHANI ENGINEERING COLLEGE

SAMPLE ADMISSION FORM

Personal Details

First name:	PARTH
Middle name:	RAJEEV
Surname:	SHARMA
Contact no: +91	9214567803

Exam Details

Physics (out of 50):	49
Chemistry (out of 50):	32
Mathematics (out of 100):	98

Research Achievements
List any research publications for which you are an author or co-author. Include any thesis or dissertation you might have completed, and any journal or conference papers (be sure to indicate the publication status, whether in preparation, submitted, accepted, or published). Do not include any papers done for a course or published in a local university publication unless it was also submitted to a national conference or journal. If the publication was not in English, indicate the language of publication.
(Enter text below)

NA

P. G. Kher Marg, (32nd Road), TPS-III Off Linking Road. Bandra (West), Mumbai - 400 050.

Figure 1 Sample Input

3.2 Data Extraction

The form will now be aligned with the template stored in the system. The system already has preprocessed coordinates of the details that need to be cropped and extracted. In this process, the handwritten data is matched to its corresponding field. The stages for data extraction are as explained below:

3.2.1 Data Segmentation

Data segmentation process crops a part of the form image to match it to its particular field. The process occurs after the image undergoes template processing. The processed image contains only handwritten data. This handwritten data image corresponding to its data field is stored for further stages of classification as seen in figure 2.

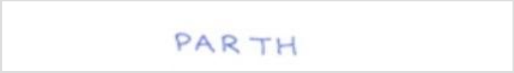


Figure 2 Segment Data Row

3.3 Preprocessing

The preprocessing [3] consists of Binarization (thresholding), Smoothing & Noise removal, Segmentation and Normalization. Preprocessing stage involves all the operations to produce a clean character image to provide an efficient input to the feature extraction stage. Before extracting features from an image, simple and common preprocessing methods are applied to systemize the data and make it feasible for the recognition algorithms. This would reduce the complexity to a certain extent.

Take an example of a college admission form. The form has fields to capture information like full name, parent or guardian's name, contact number and marks obtained in various subjects.

3.3.1 Binarization (Thresholding)

Binarization [4] is one of the important techniques for preprocessing. It transforms a colored or grayscale image into a black and white image. The Otsu's binarization technique [5] is considered as a very common and efficient technique. If a pixel value is greater than or equal to the threshold intensity, it is considered as a white ("0"). If a pixel in the image has intensity less than the threshold value, the resulting pixel is black ("1").

If an image is a color image, it is first converted into a grayscale image. The grayscale image then undergoes the process of binarization. A threshold value for each image can be obtained using MATLAB's graythresh (image) function by passing the image as a parameter. The threshold is then used to binaries the image. Figure shows the conversion of a color image to a grayscale image. Figure 3 corresponds to an image before binarization and figure 4 represents the image after the binarization process.

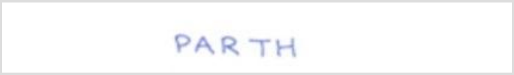


Figure 3 Image before Binarization

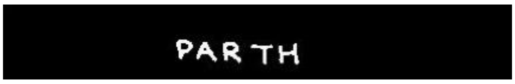


Figure 4 Image after Binarization

3.3.2 Smoothing and Noise Removal

The process of scanning introduces noise. Smoothing is used to eliminate the noises. Smoothing and noise removal can be done by filtering. Circular Mean filtering [6] is an easy to implement method of smoothing images. It reduces the amount of intensity variation between one pixel and the next.

3.3.3 Character Segmentation

Segmentation is a crucial step of OCR systems as it extracts meaningful regions for analysis. Segmentation is the process of detecting points. This method is important for recognition because it determines the start and end of a character for the correct detection of the character. This in turn affects the accuracy of the algorithm in question. Segmentation will provide the individual characters which can then be further processed for identification. Figure 5 represents the segmented characters of the input image.

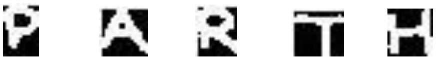


Figure 5 Character Segmentation

3.3.4 Normalization

Handwritten characters can have different sizes, positions and orientation since each human being has a unique style of writing. Hence, the character needs to be normalized and scaled as per a standard position. Normalization reduces the variation and deviation of the shapes of the characters/digits. This facilitates the feature extraction process and also improves its classification accuracy. The handwritten character recognition systems generally cope with pattern variations and distortions by linear or nonlinear pattern normalization.

3.4 Multi-class Support Vector Machine Classification

Multi-class SVM[7][8][9] aims to assign labels to instances, where the labels are drawn from a finite set of several elements. Here the set of elements are alphabets and numbers. For multi-class SVM Classification there are two classification techniques as one-versus-one and one-versus-all. A one-versus-all strategy has been implemented as it is computationally cheap and speeds up the process. Classification of instances for the one-versus-all case is done by a winner-takes-all strategy. This approach represents the earliest and most common SVM multi-class approach. Assume that there are N different classes. One versus all will train one classifier per class in total N classifiers. The aim is to construct a function that accurately predicts the class to which it belongs to.

SVMs can efficiently perform a non-linear classification using the kernel trick. The kernel trick implicitly maps their inputs into high-dimensional feature spaces. They use kernel functions to enable them to operate in a high-dimensional, implicit feature space. They do not compute the coordinates of the data in that space, but simply the inner products between the matrices of images of all pairs of data in the feature space. This operation is computationally cheaper than the explicit one. The Kernel equation used in the process is as shown in equation 1.

$$K(x,y) = \sum_{i=1}^{n} (x_{m,i}^{T} * x + 1)^{2} * \alpha_{m,i} * y_m \qquad \ldots (1)$$

Where n is number of samples, m is the character under test, and x and y are the matrices of the images.

3.5 Storage

As the data is obtained after classification it is stored locally in a spreadsheet application. In this process the classified data is matched to its corresponding field and stored in a Key-value coding compliant matrix. This spreadsheet is then used to update the server database.

4. PERFORMANCE EXPERIMENTS

Performance experiments were carried out using MATLAB (R2013A). The hardware configurations include an Intel i3 2.4 GHz processor. A Windows 32 bit operating system was used. The SVM training included data sets of different size and contents. Table 1 shows the experimental details of the digitization process. Figure 6 shows a graphical representation of number of samples versus recognition rate for different data sets.

Table 1 Performance Analysis

Dataset	Total samples	Recognition rate (%)	Time for training
Numeric	500	80	10 minutes
Alphabets	500	74	30 minutes
Alpha-Numeric	500	65	50 minutes
Numeric	1000	90	30 minutes
Alphabets	1000	78	1.5 hours
Alpha-Numeric	1000	68	2.5 hours
Numeric	2255	97	1.5 hours
Alphabets	2255	87	6 hours
Alpha-Numeric	2255	72	8 hours
Alpha-Numeric	3000	76	14 hours
Alpha-Numeric	4500	83	22 hours

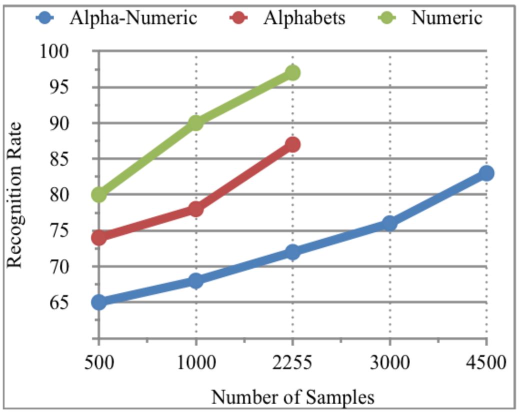


Figure 6 Graphical display of the Number of Samples versus Recognition Rate for different sample sets

5. CONCLUSION

The Handwritten Manuscript Digitizer has been implemented using the Kernel method. Kernel method is an SVM technique. The system was tested with alphabets, numbers and alpha numeric data. Numeric data consisted of 2255 samples and 97% accuracy was obtained, 87% accuracy was obtained with 2255 samples of alphabets and 84% accuracy was obtained using 4255 samples of alphanumeric data. Implementation of the system is done using MATLAB (R2013A). The system configurations include an Intel i3 2.4 GHz processor and Windows 32 bit operating system. The average time required to process a page is about 30-45 seconds.

The system can be further trained using a large number of samples to improve the accuracy. With an increased accuracy, it can be stationed at various organizations to eliminate offline data collection processes. Educational institutes can imbibe this process for procuring data in admission processes. Government organizations collect enormous data for innumerable causes like medical forms, applications for legal documents, etc. These processes can be streamlined by using the digitizer. The probability of error and time consumed will be reduced by a considerable proportion with the implementation of the system. A possible limitation for the system is cursive letters. The segmentation process for cursive letters is an arduous activity and involves over lapping of characters. The experiment was restricted to the use of capital alphabets. The use of capital alphabets resulted in satisfactory results with a noteworthy accuracy.

Enhancements can be made in the system to deploy it for the purposes of paper correction and updating marks. The digitizer can be implemented to speed up the paper correction process. Objective examination paper correction is an eminent platform for utilizing the Handwritten Manuscript Digitizer. The paper will act as input to the system and mapping the answers with the ones stored in the database. Marks can be allotted in the paper along with an entry in the system database thereby optimizing the process and reducing human efforts. The total marks scored can then be digitally stored for future evaluations.

6. REFERENCES

[1] Nasien, Dewi, Habibollah Haron, and Siti Sophiayati Yuhaniz, The Study of Handwriting Character Recognition (HCR) and Support Vector Machine (SVM), (439-447)

[2] Fabien Lauer, Ching Y. Suen, G´erard Bloch. A trainable feature extractor for handwritten digit recognition. Pattern Recognition, Elsevier, 2007, 40 (6), pp.1816-1824. <10.1016/j.patcog.2006.10.011>

[3] Character Recognition Using Matlab's Neural Network Toolbox Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre International Journal of u- and e- Service, Science and Technology Vol. 6, No. 1, February, 2013(13-20)

[4] OCR binarization and image pre-processing for searching historical documents Maya R. Gupta*, Nathaniel P. Jacobson, Eric K. Garcia 40 (2007) 389 – 397

[5] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Systems Man Cybernet Vol. 1, January 1979 (62-66).

[6] M. H. Shakoor and F. Tajeripour, Circular Mean Filtering For Textures Noise Reduction, Iranian Journal of Electrical & Electronic Engineering, Vol. 11, No. 3, Sep. 2015

[7] Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction Muhammad Naeem Ayyaz 1, Imran Javed 2 and Waqar Mahmood 3 Pak. J. Engg. & Appl. Sci. Vol. 10, Jan., 2012 (57-67)

[8] Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines John C. Platt Microsoft Research jplatt@microsoft.com Technical Report MSR-TR-98-14 April 21, 1998

[9] Shubhangi D. C., Prof. P.S. Hiremath, Handwritten English Character And Digit Recognition Using Multiclass SVM Classifier And Using Structural Micro Features, International Journal of Recent Trends in Engineering, Vol 2, No. 2, November 2009, (193-195)