Handwriting recognition using Support Vector Machine

Rishav[a] , Akash Singh[b]

[a] *Galgotias University, Greater Noida, rishav_.scsebtech@galgotiasuniversity.edu.in, INDIA*
[b] *Galgotias University, Greater Noida, akashnirban093@gmail.com, INDIA*

Abstract

Handwriting recognition contains an important part of image processing. Handwriting recognition comes under the domain of pattern recognition and the technology of artificial intelligence is extensively used in the training of the system. In handwriting, the recognition machine performs operations to automatically detect the characters and patterns written by the stoner. It is the ability of any machine to be able to distinguish individual handwritten characters using any of the supervised learning algorithms. Handwriting recognition involves various steps which includes pre – processing phase, skew detection, classification of characters and at last converting it to digital format. There is various handwriting recognition system developed in the past which are used for handwriting recognition for various character of regional languages which has an average accuracy of 95% in total. Our aim is to develop a system at university level that will be able to recognize handwritten character uploaded by students in online examinations. There is significantly less system developed for handwritten mathematical symbols recognition, thus our system will also include recognition of most of the mathematics symbols thus can be used for recognition of mathematical equation written by student in the mathematics examination. Our aim is to achieve 99% accuracy for handwritten alphanumeric characters and approx. 90% accuracy for mathematical symbols.

Keywords

Support vector machine, Convolutional neural network, Personal digital assistant, wireless local area network, Graphical user interface, Application programming interface, Model-View-Controller.

1. Introduction

Written text or symbols is one of most common means of communication. Writing of symbols text etc. is coming from the stone age itself which was used to express views or some meaningful information. Nowadays handwritten is used using pen and paper. After the verbal form of communication written form of communication is most often used to share some message, views, information etc. Handwriting is also used for personal benefits such as noting down some useful information so that one does not forget it or for writing reminders or expressing views etc. There is a lot of reason and fields for which handwritten text is used. Handwriting is very much related to physical processes because it depends on varied physical aspects of a private reminiscent of fatigue level, age, muscles, associated skeleton controlled by the brain it conjointly depends on the individual' mood whereas writing the text. Thus, there are various parameters that influence an individual's handwriting for a given population no 2 humans can have a similar form of handwriting. The handwriting of a person also depends on how his brain is conditioned or trained during growing up. The handwritten document continues to be used to record information in our day-to-day life despite the innovation of various technologies in this field . There are various technologies in the market which can be used instead of handwritten documents one major alternative is printed text using computers which are gaining popularity since the internet and technology have become more feasible thus accessible but still, the handwritten text will continue to persist for coming years as it is the easiest way of communication. Thus, our software will be useful for the masses for many coming years. Use Handwritten documents is widely in offices and schools or colleges and is often considered the best and reliable way to record and store information. Handwritten answer scripts are one of the most commonly used forms of communication in schools and colleges and

it been in use for many years, even online exams which are taking place nowadays have an option for uploading handwritten answers as students are more comfortable with manually writing the answers rather than typing it on the computer as it comes in a habit for a student since preschool they are taught to write on pages whether their views or answer for a question. Our handwriting recognition system will be based on a support vector machine algorithm which is a supervised machine learning algorithm. This algorithm is based on statistical learning theory [1].

2. Literature Review

This is a standalone section of the literature review for Our Project, the focus of which is on the study of previous research based on the Working idea of our Application. Following are the research that happened based on the application or functioning of our application.

The first significant effort in the field of character recognition a study conducted by Grimsdale in 1959. The origin of too much research work in the early sixties was based on a particular practice known as the analysis-by-synthesis method proposed by Eden in 1968. The great significance of the work of Eden was that it has been officially confirmed that all handwritten letters are composed of a limited number of scheme features, a point that was present fully incorporated into past works [2].

In 1991 the apple computer company has launched a compact device called NEWTON that is considered as significant use of handwriting recognition interface by using a small stylus pen. User can make use of this pen to write on newton screen that can convert handwritten text into digital text. The developed system was less than the ideal conditions and its acceptance among crowd was never great. [3]. In 2000's palm company has developed another handwriting recognition system called graffiti which had become mainstream in offices uses but it's accuracy was being questioned and thus the model become extinct soon [4]. Microsoft corporation has also made an attempt in this field by launching a tablet for handwriting recognition, it was not dependent on user's new answers, but it was extensively trained on labelled dataset. This system was supposed to have higher success rate than the learning system for most of the users, but its reliability was in question. Development of handwriting recognition has become mainstream by the introduction of cellular devices and personal digital assistant (PDA) and companies are investing multi-billion dollars in this technology so as to develop a highly reliable and accurate handwriting recognition system[5].

In this paper the system was developed using the technologies support vector machine and a significant use of dynamic time wrapping (DTW) by developing a SVM kernel. The main advantage of the system discussed in this paper was that it creates discriminant class boundaries which is not sensitive to initial model presumptions [6]. This paper proposes a system based on the combination of K-nearest neighbor (KNN) and support vector machine (SVM) supervised learning algorithms for extensive handwriting recognition. The KNN algorithm is used for identifying classes of characters and SVM is used for detection of similar classes for each individual characters [7]. This paper proposes a learning model for handwriting recognition that is based on convolutional neural network (CNN). This paper compares the accuracy of the models developed using support vector machine and convolutional neural network gives similar results in terms of both accuracy rate and the processing time. The system developed gives an accuracy rate of 93.05% for uppercase alphabets, 86.21% for lowercase alphabets and 91.37 for numeric characters [8]. This paper discusses a system that uses a feature extraction method for the recognition of handwriting of Gurumukhi characters. This technique was based on the boundary extent of the text image and used different types of feature selection techniques to reduce the dimension of feature vectors [9].

3. Problem Formulation

Handwriting recognition is the ability of a machine to reproduce human handwriting in digital form. In addition to recognition, handwriting analysis is often used to identify a person's writing, compare two handwritten samples, or analysis to extract other information such as human characteristics or even to identify potential diseases. Handwriting text recognition has a variety of applications such as signature verification for the authorization of a person, to convert handwritten text into digital form so that everyone can read it free from abnormalities that come in an individual's handwriting. During covid times all universities and schools were conducting online examinations some were providing options to students to type the answer, upload handwritten answer scripts or both. Since students were having a habit of writing the answer on the page so it was new for them to type on their PC, as it was new for them, and many students were not completely comfortable with this as typing on an electronic device needs practice and exam there were time constraints, so it is observed many students have opted for writing the answer and then uploading it method which is ok. But the main problem is that our software will tackle is that difficulties faced by teachers while checking the answer of the students that they have uploaded as the in some cases answer sheets were not properly visibly due to scanning in improper lighting and quality degradation due to restrictions in the size of the file to be uploaded. To tackle these problems our software will integrate with LMS or whatever exam conducting software an institution has and provide appropriate text for given handwriting. Our software will provide two benefits first which was stated as above handwriting to text conversion second, we will also be able to verify that the answer uploaded by the student was written by himself/herself or not to eliminate unfair practices. There is various handwriting recognition software out there in the market but there is still not a single perfect handwriting recognition software which is released or is in use as some of them have a low accurate result or some are expensive to own, to eliminate these situations we are designing a handwriting recognition software which will be based on supervised machine learning algorithm which is support vector machine.

4. Required Tools.

Our handwriting recognition system will be based on a supervised machine learning algorithm support vector machine (SVMs, also support vector networks) therefore a detailed knowledge that how this algorithm works will be essential. It is used for classification and regression problems in Machine Learning. The main purpose of the Support vector machine classifier algorithm is to construct the single line or decision boundary which will divide n-dimensional space into classes in order that we will easily put the new datum within the correct category in the future. This best decision boundary is named a hyperplane [10].

SVM chooses the acute points/vectors that help in creating the hyperplane. The extreme cases which decide constraints of the width of the hyperplane are called support vectors, and hence algorithm is termed a Support Vector Machine.

For the purpose of training our model and subsequently testing it for accuracy a publicly available database will be required and for the next phase of the database we will be requiring samples of handwritten documents from students at our university so that our proposed system will be able to solve its purpose.

For the development of the system, we will be requiring a PC with, an IDE installed to write the python code. For the running of our python code, we will additionally be requiring a python compiler installed in our system or as an extension with the IDE we will be using. Physical devices that will be required during the development phase are a scanner for scanning the handwritten samples connected and a physical storage device as these handwritten samples can be in thousands and thus the size of the scanner images can go in terabyte [11].

5. Project Design

5.1 Database

For training and better handwriting recognizer mechanism a standard database will be required. In the initial phase of testing and training we will use MNIST handwritten digit database to train our model and test subsequently for performance. The MNSIT database of handwritten digits, is publicly available dataset of handwritten digits, contains a training set of 60,000 examples, and a test set of 10,000 examples [MNSIT Kaggle dataset 2017(version 6.0)]. The digits have been size-normalized and cantered in a fixed-size image. The dataset consists of 8X8 pixel images of digit. The image attribute of the dataset stores 8X8 arrays of grayscale values for each image After initial testing is over, we will create our own database in which handwritten samples of the college students will be recorded and captured and will be stored as an image file. This will allow us to train our model on a large scale thereby increasing the accuracy and testing for performance can also be carried out afterwards. At later stage of our software development process once the system is trained and tested using MNSIT database, then we will be training and testing our system using real world test sample collected from a group of students studying in different colleges of
Greater Noida, so that our system can effectively developed for real case scenario. The handwritten sample for database of 10,000 English characters will be taken from around 312 students irrespective of their age and gender. Upon collecting the sample, the process for training and testing of the system will be same as it is for MNSIT database.

5.2 Training

The aim of this project was to achieve 99% efficiency of the system in individually identifying each character. Thus, after initial training and testing of the system through MNSIT database, the system was trained on 15 distinguished sample of each individual character of in English alphabets and also numeric characters and perform subsequent testing of the trained model.

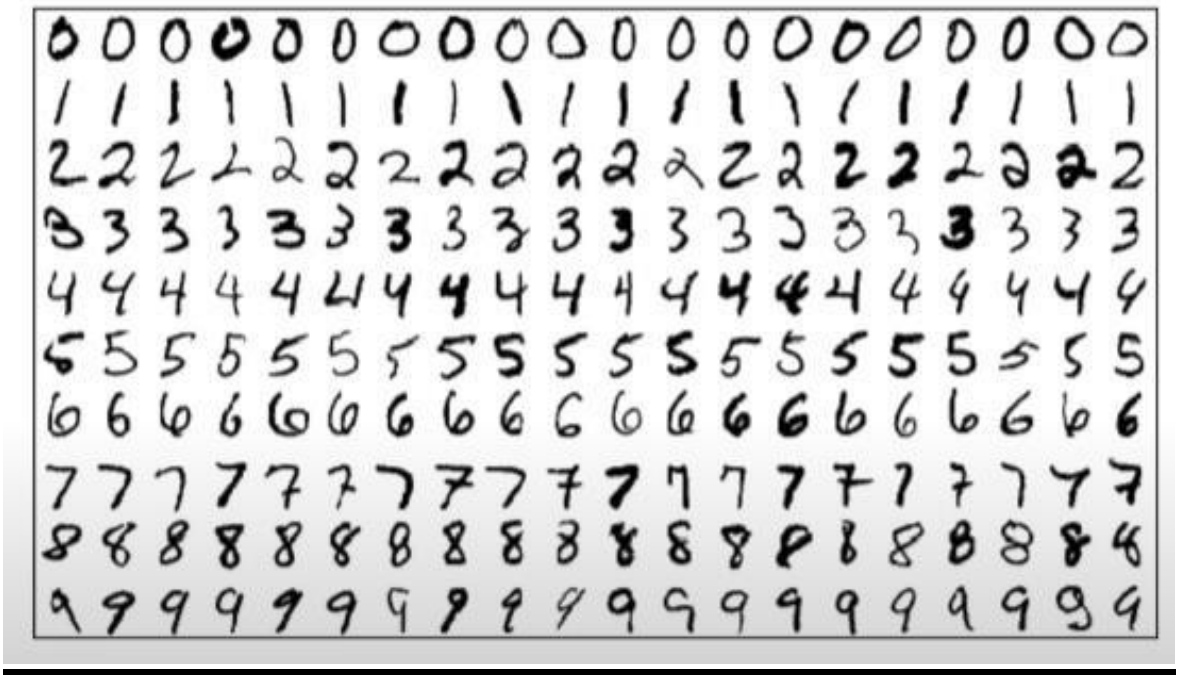

Fig1 MNSIT Database [Kaggle dataset version 6.0]

5.3 Pre – processing Phase

For training and a better handwriting recognizer mechanism, a standard database will be required. Since our application will serve its purpose at university level or to universities therefore data (handwritten characters) handwritten sample was collected from students at different universities located in greater Noida, Uttar Pradesh without imposing any kind of constraints. The collected sample represents a vast variety of writing fonts. Scanning of these samples is carried out by a flat- bed scanner.

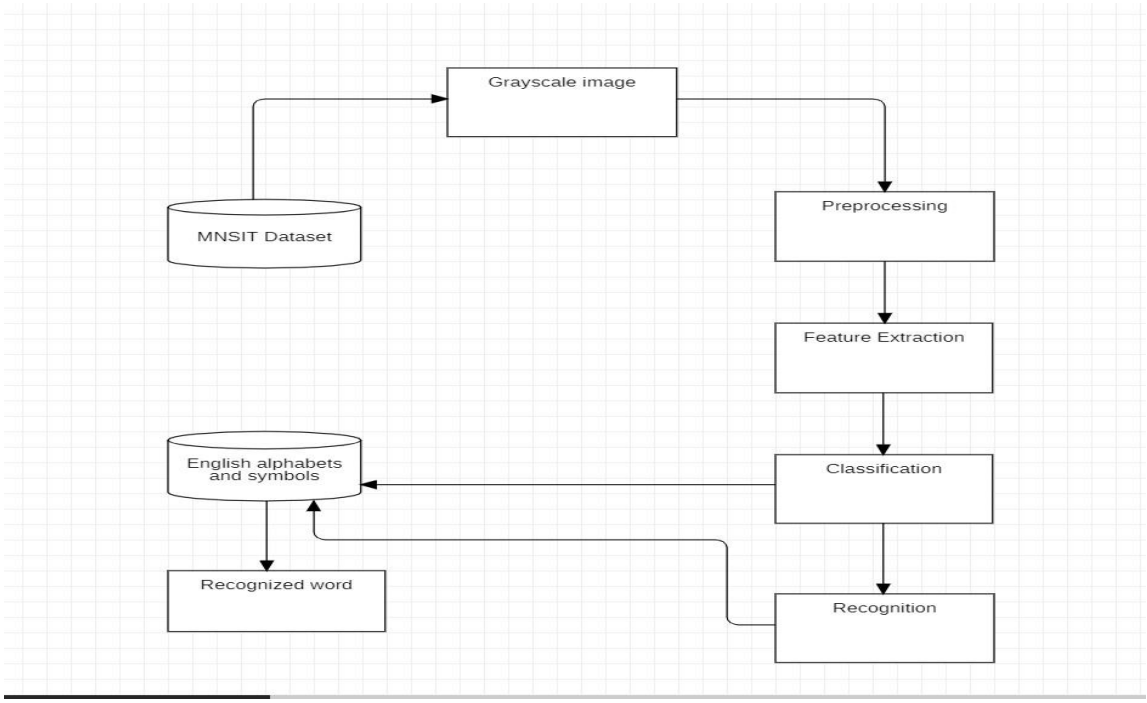

Fig 2. Flowchart diagram for the proposed system

5.3.1 Removal of noise from image

Noise is defined as the variation or derivation of brightness or color information in the images. Noise in an image is often referred to as digital noise. The source of the noise in the image is mostly due to the following conditions such as insufficient lighting, Environmental conditions, sensor temperature, transmission channels and dust factor. A median filter will be used to remove this unwanted noise from the image scanned. It is a digital filtering technique that is typically used during the pre-processing phase.

5.3.2. Skew Detection

The line which is perfectly horizontal has an angle of 90 degrees in its row theta space. Scanning of images for database procurement using a scanner may have a skew in the image scanned. It is measured with respect to page borders and can be corrected by simply rotating the image. But for better accuracy, we can use Hough transform method. Hough transform works as to suppose there is x and y plane, and our character lies along the line with equation y=ax+b (described in terms of x and y). Here parameters a and b are used to define the angulation of the line. The Hough transform method converts the image to polar coordinates. This process finds coordinates on the given line and a combination of coordinates are oriented if disorientation is found of about +-15 degree then it rotates the set of coordinates to have proper horizontal orientation w.r.t page border [12].

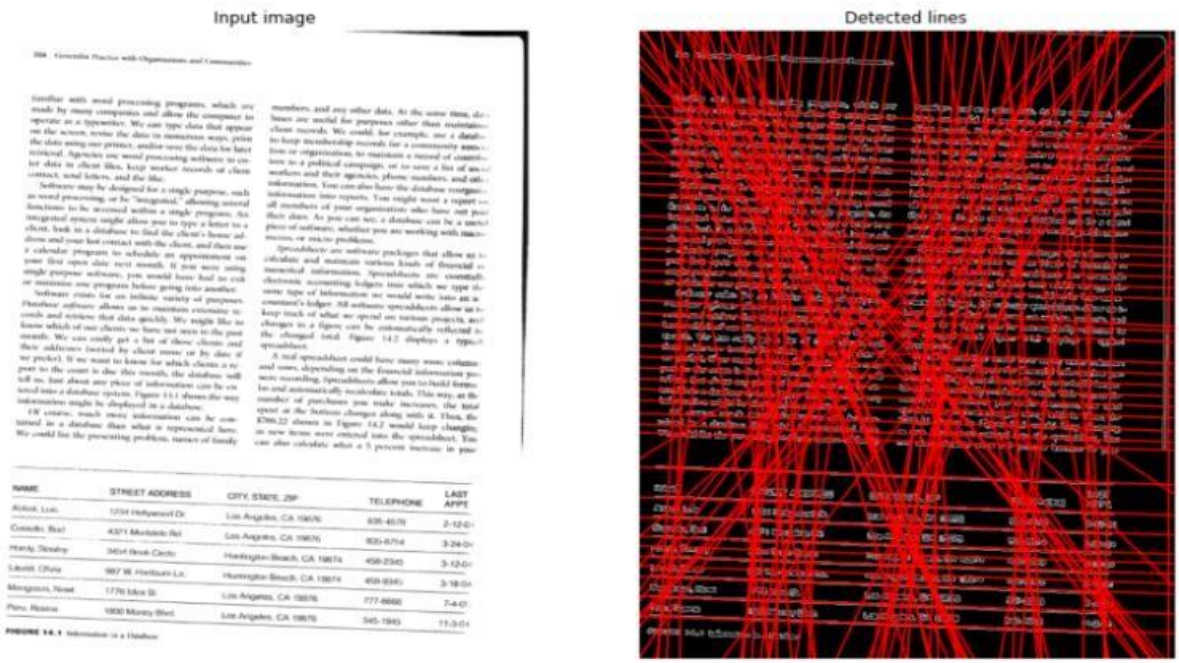

Fig 3. Hough transformation of skew lines

5.3.3. Segmentation

The profile of the horizontal histogram, i.e., an effective number of black pixels each row, included a low value between the highest values each line is used separation. From each line containing characters, letter segregation first by labelling the connected object. A small bound rectangle containing part is extracted and stored in the database. Contains broken and distorted characters too.

5.3.4. Feature Extraction

Each character in each image is represented by a vector of features. These features are computed as follows: A, where a is the pre – processed character image and B is the derived global feature.

5.4 Support Vector Machine

Support vector machine is a supervised learning technique used for linear and non – linear classification. It has comparatively great performance from other supervised learning technique and is base4d on statistical learning theory.

The SVM Classifier was proposed by Vapnik in 1995 for solving both classification and regression problems but in majority it is used for solving classification problems. SVM is used for binary classification and the learning algorithm behind it comes from a separating hyper-plane is binary classification, a linear decision function y=f(x) is used where $f(x) = w^T a + x$ where w is the weight vector and x is a bias[13].

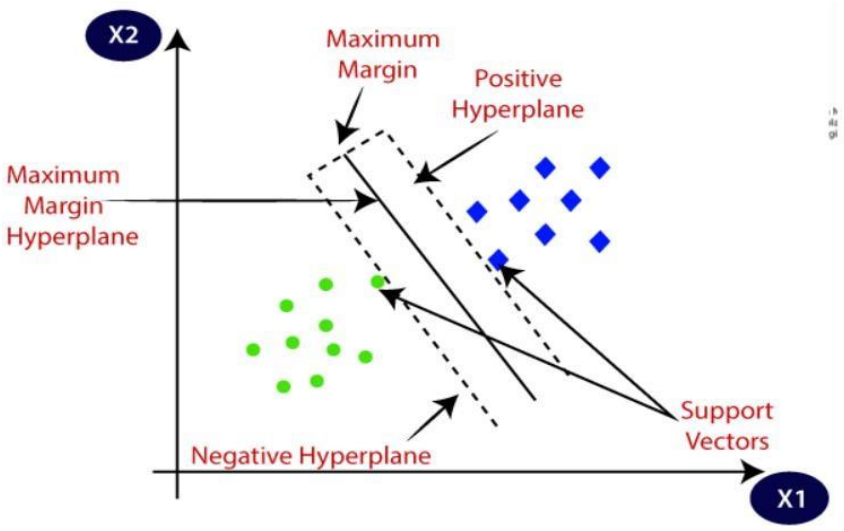

Fig 4. Support vector machine execution diagram

Classification is determined by the sign of linear decision function f(x) as it could be +1 or -1. The solution obtained will be optimal if the hyper plane is optimal if the hyper plane is placed between the two classes (fig 4). The points that decide the width of hyper plane are called support vectors. Hence this technique is known as support vector machine. The separation between the two classes can be given by 2/11W for linearly inseparable data; it is nearly impossible to find out linear decision function, so the initial pattern space is arranged into a high dimensional future space using some the nonlinear mapping functions. This process allows to map a nonlinear problem from low level separating the classes can be obtained. For a training set of case – label pair (U_i,V_i), i=1,2,3,4…….,l. where U € R_n and V € [+1,-1]. We require the solution of the following optimal problem [14].

$$\text{Min}_{w,\,x,\,e} \quad 1/2\; w^T w + c\sum_{i=0}^{l} e(i)$$

Subject to

$$(v_i\, w^T \alpha(U_i) + x) \geq 1 - e_i$$

$$e_i \geq 0,\ i = 1,2,\ldots\ldots,\ l$$

The training vector Z_i are mapped into a higher dimensional space using function α, the support vector machine classifier gives a linear separating hyper plane with highest margin in the given mapped space. Practically it is not possible to construct a hard margined hyper plane because some classes may overlap each other due to noise. The e_i is given as a slack variable to modify the constraints and x is the bias term. The Kernel function is termed as

$$K_f(U_i,\ U_j) = \alpha(U_i)^T \ .\ \alpha(U_j)$$

Basic kernel functions include:

$$\text{Linear: } K_f\,(U_i, U_j) = u_i^T \,.\, u_j$$

$$\text{Polynomial: } K_f\,(U_i, U_j) = (\, u_i^T \,.\, u_j + 1)^d$$

$$\text{Radial basis function: } K_f\,(U_i, U_j) = \exp\,(\, -a \,||\, u_i - u_j \,||^2\,),\ \mu > 0$$

$$\text{Sigmoid: } K_f(U_i, U_j) = \tanh(au_i^T . u_j + r)$$

Where μ, r and d are kernel parameters. The functioning of support vector machine depends on the selection of kernel, the kernel's parameter, and soft margin parameter C.

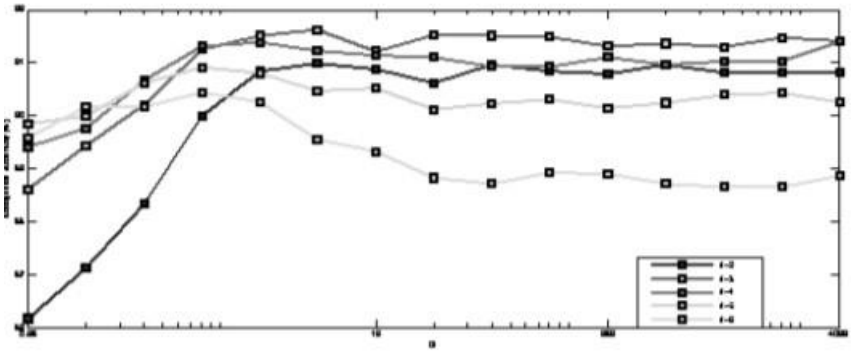

Fig 5. Execution plot for polynomial kernel for d= [2, 3, 4, 5, 8]

6. Result and Conclusion

The experiment had been performed on a database of 10,000 handwritten English characters and numeric characters written by 312 different writers. It contains all the 62 basic characters for representing each alphabet and number as a vector of features, we have used grid of different sizes i.e., 16x16, 8x8, 4x4, and 2x2 were put down on the pre – processed image, giving the 4, 16, 64 and 256 features because of occurrence of pixels in each ell. Along with these features, 4 global features which are number of endpoints, width/height ratio, number of branch points, and number of cross points are included. The number of branch points, number of endpoints, and number of cross points are calculated from a thinned image whereas width / height ratio is calculated from initial segmented images. All of this represents the feature sets having 9, 20, 68, and 260 feature.

Class	char	Error rate	class	char	Error rate	class	char	Error rate	class	char	Error rate
1	A	.0045	17	Q	.001	33	g	.0045	49	w	.0045
2	B	.0025	18	R	.0025	34	h	.0055	50	x	.0035
3	C	.004	19	S	.0025	35	i	.0045	51	y	.003
4	D	.009	20	T	.0055	36	j	.003	52	z	.003
5	E	.0015	21	U	.005	37	k	.0025	53	0	.001
6	F	.0065	22	V	.0045	38	l	.002	54	1	.001
7	G	.0055	23	W	.0015	39	m	.0055	55	2	.0025
8	H	.0055	24	X	.003	40	n	.0045	56	3	.0055
9	I	.003	25	Y	.0035	41	o	.001	57	4	.0055
10	J	.0015	26	Z	.0025	42	p	.0025	58	5	.0045
11	K	.0045	27	a	.0035	43	q	.0025	59	6	.004
12	L	.002	28	b	.0015	44	r	.0015	60	7	.0025
13	M	.0035	29	c	.001	45	s	.003	61	8	.0055
14	N	.001	30	d	.001	46	t	.0025	62	9	.005
15	O	.0015	31	e	.0045	47	u	.0035			
16	P	.004	32	f	.0045	48	v	.002			

Table1: Error rate for every independent character using RBF kernel for 13 samples taken.

Training sets and tested using the remaining bottom set. Training is repeated 10 times and the total amount of recognition all 10 subsets are not included in it training data are calculated. Tests they have done with 8, 20, 68 feature sets and 260 features. In all tests, I A database is categorized by a random process into training (80%) and testing (20%). From the first phase itself was trivial that the one who offers the best performance was based on a 4 × 4 grid (68 features). In this paper, the idea of seeing offline manuscripts successfully offline has become proposed. It is found to be simple but effective compared to the existing handwriting character recognition. It is possible to integrate this advanced application into existing applications that require character acquisition. Various programs can use this model bank check, which helps to restore text in the image, text recognition from business cards, to help the blind see handwritten text. Efficiency has increase due to the separation of the edge between the hyperplane planes. However, our application designed cannot process handwritten characters. In the future, support for find and see compound characters can be added.

7. References

[1] Vapnik, V.N., 2019. An overview of statistical learning theory. *IEEE transactions on neural networks*, *10*(5), pp.988-999.

[2] Mantas, J., 2015. An overview of character recognition methodologies. *Pattern recognition*, *19*(6), pp.425-430.

[3] Roberts, G.I., and Samuels, M.T., 2018. Handwriting remediation: A comparison of computerbased and traditional approaches. *The Journal of Educational Research*, *87*(2), pp.118-125.

[4] Chang, C.J., Lo, C.O. and Chuang, S.C., 2020. Applying Video Modeling to Promote the Handwriting Accuracy of Students with Low Vision Using Mobile Technology. *Journal of Visual Impairment & Blindness*, *114*(5), pp.406-420.

[5] Potanin, M., Dimitrov, D., Shonenkov, A., Bataev, V., Karachev, D. and Novopoltsev, M., 2021. Digital Peter: Dataset, Competition and Handwriting Recognition Methods. *arXiv preprint arXiv:2103.09354*.

[6] [Bahlmann, C., Haasdonk, B. and Burkhardt, H., 2012, August. Online handwriting recognition with support vector machines-a kernel approach. In Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition (pp. 49-54). IEEE.]

[7] Zanchettin, C., Bezerra, B.L.D. and Azevedo, W.W., 2020, June. A KNN-SVM hybrid model for cursive handwriting recognition. In The 2012 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[8] Fanany, M.I., 2019, May. Handwriting recognition on form document using convolutional neural network and support vector machines (CNN-SVM). In 2017 5th international conference on information and communication technology (ICoIC7) (pp. 1-6). IEEE.

[9] Kaur, R.P., Jindal, M.K. and Kumar, M., 2021. Text and graphics segmentation of newspapers printed in Gurmukhi script: a hybrid approach. *The Visual Computer*, *37*(7), pp.1637-1659.

[10] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A., 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, pp.189-215.

[11] Ahmad, A.R., Khalia, M., Viard-Gaudin, C. and Poisson, E., 2004, November. Online handwriting recognition using support vector machine. In *2004 IEEE Region 10 Conference TENCON 2004.* (pp. 311-314). IEEE.

[12] Kunte, R.S. and Samuel, R.S., 2007. A simple and efficient optical character recognition system for basic symbols in printed Kannada text. Sadhana, 32(5), p.521.

[13] Shah, M. and Jethava, G.B., 2013. A literature review on handwritten character recognition.

[14] Kumar, M., Sharma, R.K. and Jindal, M.K., 2011. SVM based offline handwritten Gurmukhi character recognition. SCAKD Proceedings, 758, pp.51-62.