Optical Character Recognition

Deepali D. Kamat

Abstract—**The process of detecting and translating hand-written, typed or printed text into a digital format is known as Optical Character Recognition (OCR). OCR is performed with the intent of converting data at hand into a format comprehended by machines in-order to perform operations like searching, indexing and editing. The steps for OCR involve reading text from paper in the form of images or scanned documents and converting them into a machine editable form (for example, ASCII codes) such that operations like indexing, searching and editing may be performed. The use of OCR permits us to convert the data in a book or a magazine article; obtained and stored in image or PDF format, and edit it using a text editor or a document reader. The OCR system comprises of an optical scanner that helps in reading text, and software to analyze the input. OCR systems can be designed to detect text in large variety of fonts and even handwritten texts, at advanced levels. OCR systems have enormous potential as they help users to harness the power of computers by accessing printed documents in digital format. The OCR systems are being used widely in legal professions, where document searches can be performed in a time span of few seconds instead of days, when done manually. This paper describes how Support Vector Machine(SVM) may be used to perform the task of character recognition.**

Index Terms—**data preprocessing, support vector machines**

I. INTRODUCTION

A major amount of information in this world is stored in the form of hard copy documents. There is a need to convert the printed, typed or handwritten information into a digital storage format for safe keeping of the information, easier access and effective storage. OCR is a process which provides alphanumeric recognition of handwritten and printed characters electronically by scanning or image acquisition. The recognition system interprets the images into ASCII data. This helps track every evident piece of information being processed while making human and computer interaction simpler by manifold. Character recognition systems are of two types, offline handwritten text recognition and online handwritten text recognition. In offline handwritten text recognition system, information or data written on paper are scanned and saved as an image for further processing where as in online handwritten recognition systems, the inputs given to digital devices like tablets using stylus are taken in as two dimensional coordinates of successive points which are represented as a function of time and of the order of strokes made by the writer. [8] OCRs are proving to be an indispensable part of document scanners and are also used in various applications like postal processing, writer identification, license plate recognition system, smart card processing system, automatic data entry, bank check processing, postal automation, address and zip code recognition, script recognition, document reading, mail sorting, signature verification, language identification, banking, security applications like passport authentication among many other myriad applications. Many organizations are incorporating OCR systems into practice to eliminate human interactions improving efficiency and reducing error rates.

A. Phases of OCR

i. Data Acquisition:
The initial phase of the OCR system. Involves gathering images from device sensors and personal digital assistants (online handwritten) and scanned document images (offline handwritten).

ii. Preprocessing:
Image enhancement to simplify pattern recognition. Includes noise removal and improving data consistency. Preprocessing transforms the format of data to decrease variation and create effective processing in further steps. This includes eliminating all the areas that might reduce the recognition rate and increase the complexity of the problem. Preprocessing includes steps like binarization of images, noise removal, smoothing, skew rectification, normalization, thresholding etc.

iii. Segmentation:
The process of segmentation is an integral part of image preprocessing as this step

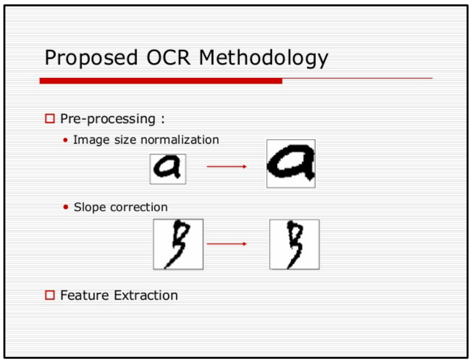
ensures efficiency and accuracy of the recognition system. Segmentation produces isolated characters which are reduced to specific sizes depending on the task at hand.

iv. Normalization:
The output of segmentation is a matrix of mxn dimensions. The output matrix is normalized by reducing the size and removing the excessive and repetitive information from the input image without eliminating the important details of the image.

v. Feature Extraction:
This step involves the selection and extraction of important features in the form of vectors. These vectors are given as inputs to classifying units where they are classified into different classes based on the feature type.

vi. Classification:
The final stage of OCR includes training the system using multiple combinations of multilayer perceptron [8] (k-Nearest Neighbor (k-NN), Bayes Classifier, Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM) etc.) The decision of the choice of the classifier depends on factors like training set, number of available parameters, etc.

vii. Post Processing:
Post processing helps in incorporating context and feature information in all the stages of the OCR. This improves the recognition rate of the character.



Proposed OCR Methodology

□ Pre-processing :
- Image size normalization
- Slope correction

□ Feature Extraction

Figure 1: Image Preprocessing

The main objective of optical character recognition is to classify the alphanumeric or other optical samples which are mostly saved as digital images. Support Vector Machines are commonly used to solve sample recognition problems. [1] SVM is the technique of creating functions using the set of labeled training data at hand. [9] These functions created could either be regression functions or classification functions. The SVM calculates a hyper-plane taking into consideration the most optimum separation among the data points belonging to various classes in the high dimensional vector space. This optimal hyper-plane obtained for the patterns help determine the solution for the optimization.

$$\frac{1}{2}w^T w + \frac{C}{2}\xi^T \xi \qquad (1)$$

Such that,

$$D(w^T \phi(x) - \gamma e) + \xi \leq e \qquad (2)$$

Where, w is the coefficient vector, γ is the bias term, C is the cost factor, ϕ(x) is the non-linear mapping function that maps input vector to higher dimensional space, ξ is the slack variable, D is the diagonal matrix containing class values.

II. BACKGROUND

The data used in this project is from the UCI Machine Learning Repository [] where the character images were based on 20 different fonts. Each letter in the 20 font batch are randomly distorted images which were distorted to produce a file of 20,000 unique stimuli. Each unique of these stimuli were converted into 16 primitive numerical attributes consisting of statistical moments and edge counts which were scaled to fit into a range of integer values from 0 through 15. We train on the first 16000 values and then use the resulting model to recognize the characters for the remaining 4000 values.

MATLAB software was used to implement the character recognition process. An open source machine learning library, LIBSVM was used to implement the character recognition process. LIBSVM implements the sequential minimal optimization algorithm for kernalized support vector machines, supporting the technique of classification and regression.[1] The values in the data set are alphabets of the English language and need to be converted into equivalent numeric values (A corresponds to 1, B to 2 and so on) as the alphabets in their original form are not compatible with the LIBSVM library.

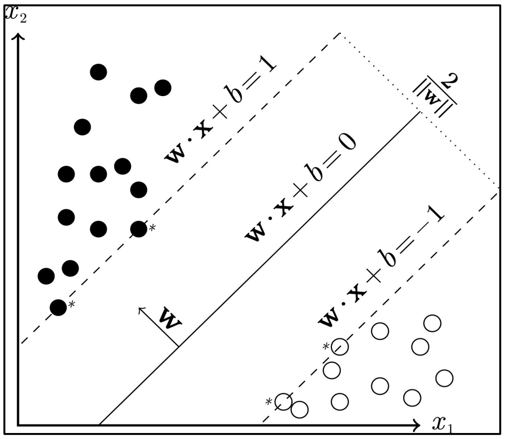

x_2
$\mathbf{w}\cdot\mathbf{x}+b=1$
$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w}\cdot\mathbf{x}+b=0$
$\mathbf{w}\cdot\mathbf{x}+b=-1$
$\mathbf{w}$
*
x_1

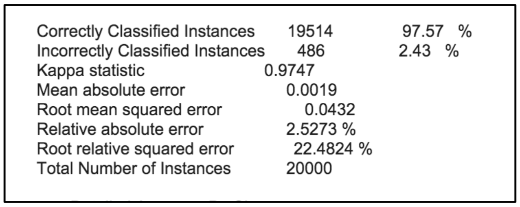
Figure 2: SVM Descriptive Plot[17]

III. IMPLEMENTATION DETAILS

The data taken from the Machine learning repository is of labeled form. This data falls in the fifth stage of the phases of OCR. The extracted labeled data is given as an input to the classification stage where an SVM is used to test the accuracy of the system. SVM efficiently perform non-linear classification using the kernel trick, by mapping their inputs into high-dimensional feature spaces.

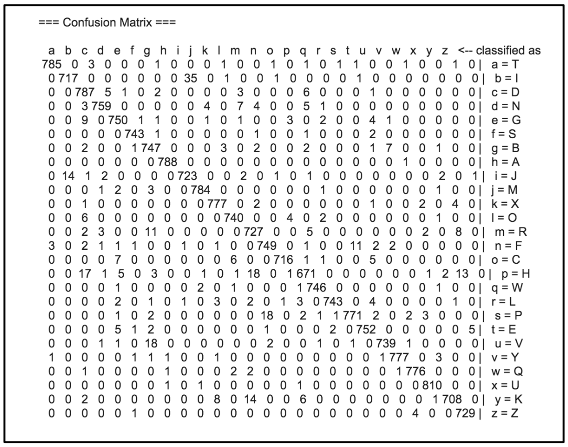
IV. RESULTS

The following results were obtained:



```
Correctly Classified Instances      19514          97.57  %
Incorrectly Classified Instances      486           2.43  %
Kappa statistic                     0.9747
Mean absolute error                  0.0019
Root mean squared error               0.0432
Relative absolute error              2.5273 %
Root relative squared error           22.4824 %
Total Number of Instances            20000
```

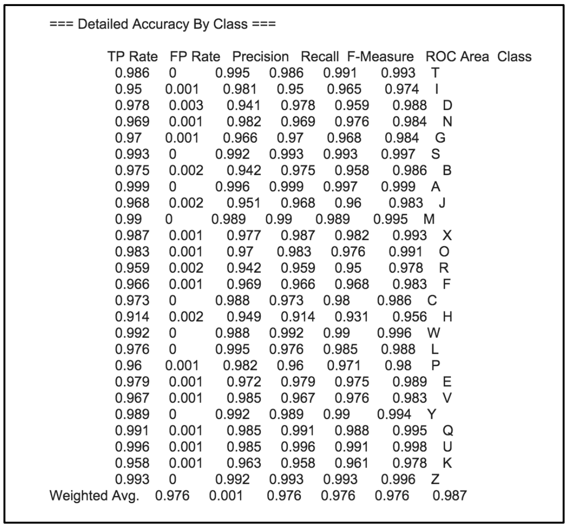
Figure 3: Result



```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   s   t   u   v   w   x   y   z   <-- classified as
 785   0   3   0   0   0   1   0   0   0   1   0   0   1   0   1   0   1   1   0   0   1   0   0   1   0 |   a = T
   0 717   0   0   0   0   0   0  35   0   1   0   0   1   0   0   0   0   1   0   0   0   0   0   0   0 |   b = I
   0   0 787   5   1   0   2   0   0   0   0   3   0   0   0   6   0   0   0   1   0   0   0   0   0   0 |   c = D
   0   0   3 759   0   0   0   0   0   4   0   7   4   0   0   5   1   0   0   0   0   0   0   0   0   0 |   d = N
   0   0   9   0 750   1   1   0   0   1   0   1   0   0   3   0   2   0   0   4   1   0   0   0   0   0 |   e = G
   0   0   0   0   0 743   1   0   0   0   0   0   1   0   0   1   0   0   0   2   0   0   0   0   0   0 |   f = S
   0   0   2   0   0   1 747   0   0   0   3   0   2   0   0   2   0   0   0   1   7   0   0   1   0   0 |   g = B
   0   0   0   0   0   0   0 788   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0 |   h = A
   0  14   1   2   0   0   0   0 723   0   0   2   0   1   0   1   0   0   0   0   0   0   0   2   0   1 |   i = J
   0   0   0   1   2   0   3   0   0 784   0   0   0   0   0   0   1   0   0   0   0   0   0   1   0   0 |   j = M
   0   0   1   0   0   0   0   0   0   0 777   0   2   0   0   0   0   0   0   1   0   0   2   0   4   0 |   k = X
   0   0   6   0   0   0   0   0   0   0   0 740   0   0   4   0   2   0   0   0   0   0   0   1   0   0 |   l = O
   0   0   2   3   0   0  11   0   0   0   0   0 727   0   0   5   0   0   0   0   0   0   2   0   8   0 |   m = R
   3   0   2   1   1   1   0   0   1   0   1   0   0 749   0   1   0   0  11   2   2   0   0   0   0   0 |   n = F
   0   0   0   0   7   0   0   0   0   0   0   6   0   0 716   1   1   0   0   5   0   0   0   0   0   0 |   o = C
   0   0  17   1   5   0   3   0   0   1   0   1  18   0   1 671   0   0   0   0   0   0   1   2  13   0 |   p = H
   0   0   0   0   1   0   0   0   0   2   0   1   0   0   0   1 746   0   0   0   0   0   0   1   0   0 |   q = W
   0   0   0   0   2   0   1   0   1   0   3   0   2   0   1   3   0 743   0   4   0   0   0   0   1   0 |   r = L
   0   0   0   0   1   0   2   0   0   0   0   0   0  18   0   2   1   1 771   2   0   2   3   0   0   0 |   s = P
   0   0   0   0   5   1   2   0   0   0   0   0   0   1   0   0   0   2   0 752   0   0   0   0   0   5 |   t = E
   0   0   0   1   1   0  18   0   0   0   0   0   0   2   0   0   1   0   1   0 739   1   0   0   0   0 |   u = V
   1   0   0   0   0   1   1   1   0   0   1   0   0   0   0   0   0   0   0   0   1 777   0   3   0   0 |   v = Y
   0   0   1   0   0   0   0   1   0   0   0   2   2   0   0   0   0   0   0   0   0   1 776   0   0   0 |   w = Q
   0   0   0   0   0   0   0   1   0   1   0   0   0   0   0   1   0   0   0   0   0   0   0 810   0   0 |   x = U
   0   0   2   0   0   0   0   0   0   0   8   0  14   0   0   6   0   0   0   0   0   0   0   1 708   0 |   y = K
   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   4   0   0 729 |   z = Z
```

Figure 4: Confusion Matrix



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.986     0          0.995      0.986     0.991      0.993    T
                 0.95      0.001      0.981      0.95      0.965      0.974    I
                 0.978     0.003      0.941      0.978     0.959      0.988    D
                 0.969     0.001      0.982      0.969     0.976      0.984    N
                 0.97      0.001      0.966      0.97      0.968      0.984    G
                 0.993     0          0.992      0.993     0.993      0.997    S
                 0.975     0.002      0.942      0.975     0.958      0.986    B
                 0.999     0          0.996      0.999     0.997      0.999    A
                 0.968     0.002      0.951      0.968     0.96       0.983    J
                 0.99      0          0.989      0.99      0.989      0.995    M
                 0.987     0.001      0.977      0.987     0.982      0.993    X
                 0.983     0.001      0.97       0.983     0.976      0.991    O
                 0.959     0.002      0.942      0.959     0.95       0.978    R
                 0.966     0.001      0.969      0.966     0.968      0.983    F
                 0.973     0          0.988      0.973     0.98       0.986    C
                 0.914     0.002      0.949      0.914     0.931      0.956    H
                 0.992     0          0.988      0.992     0.99       0.996    W
                 0.976     0          0.995      0.976     0.985      0.988    L
                 0.96      0.001      0.982      0.96      0.971      0.98     P
                 0.979     0.001      0.972      0.979     0.975      0.989    E
                 0.967     0.001      0.985      0.967     0.976      0.983    V
                 0.989     0          0.992      0.989     0.99       0.994    Y
                 0.991     0.001      0.985      0.991     0.988      0.995    Q
                 0.996     0.001      0.985      0.996     0.991      0.998    U
                 0.958     0.001      0.963      0.958     0.961      0.978    K
                 0.993     0          0.992      0.993     0.993      0.996    Z
Weighted Avg.    0.976     0.001      0.976      0.976     0.976      0.987
```

Figure 5: Accuracy Obtained

V. FUTURE WORK

The next steps of the project involve taking raw scanned images of documents and perform feature extraction using HoG (Histogram of Oriented Gradients) filters. This filter is a feature descriptor used in image processing and computer vision with the purpose of object detection. HoG filters count the occurrences of gradient orientation in each local section of an image. This method performs similar actions as that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but this varies in it's computation process on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. [] These features are extracted and then the data obtained is preprocessed so that it can be segmented and given as an input to the multilayer perceptron classification system.

REFERENCES

[1] Jianhong Xie, "Optical Character Recognition Based on Least Square Support Vector Machine," in Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on , vol.1, no., pp.626-629, 21-22 Nov. 2009

[2] Ramanathan, R.; Ponmathavan, S.; Valliappan, N.; Thaneshwaran, L.; Nair, A.S.; Soman, K.P., "Optical Character Recognition for English and Tamil Using Support Vector Machines," in Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT '09. International Conference on , vol., no., pp.610-612, 28-29 Dec. 2009

[3] Pongsametrey Sok; Nguonly Taing, "Support Vector Machine (SVM) based classifier for Khmer Printed Character-set Recognition," in Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA) , vol., no., pp.1-9, 9-12 Dec. 2014

[4] Fedorovici, L.; Dragan, F., "A comparison between a neural network and a SVM and Zernike moments based blob recognition modules," in Applied Computational Intelligence and Informatics (SACI), 2011 6th IEEE International Symposium on , vol., no., pp.253-258, 19-21 May 2011

[5] Ramanathan, R.; Soman, K.P.; Thaneshwaran, L.; Viknesh, V.; Arunkumar, T.; Yuvaraj, P., "A Novel Technique for English Font Recognition Using Support Vector Machines," in Advances in Recent Technologies in Communication and Computing, 2009. ARTCom '09. International Conference on , vol., no., pp.766-769, 27-28 Oct. 2009

[6] Kilic, N.; Gorgel, P.; Ucan, O.N.; Kala, A., "Multifont Ottoman character recognition using support vector machine," in Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on , vol., no., pp.328-333, 12-14 March 2008

[7] Wikipedia contributors. Support vector machine. Wikipedia, The Free Encyclopedia. November 17, 2015, 06:10 UTC.

[8] Wikipedia contributors. LIBSVM. Wikipedia, The Free Encyclopedia. March 2, 2015, 10:49 UTC.

[9] Wikipedia contributors. Optical character recognition. Wikipedia, The Free Encyclopedia. October 30, 2015, 17:30 UTC

[10] Arnold, R.; Miklos, P., "Character recognition using neural networks," in Computational Intelligence and Informatics (CINTI), 2010 11th International Symposium on , vol., no., pp.311-314, 18-20 Nov. 2010.

[11] Farhad, M.M.; Nafiul Hossain, S.M.; Khan, A.S.; Islam, A., "An efficient Optical Character Recognition algorithm using artificial neural network by curvature properties of characters," in Informatics, Electronics & Vision (ICIEV), 2014 International Conference on , vol., no., pp.1-5, 23-24 May 2014.

[12] Pradeep, J.; Srinivasan, E.; Himavathi, S., "Neural network based handwritten character recognition system without feature extraction," in Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on , vol., no., pp.40-44, 18-19 March 2011.

[13] Xiaojun Zhai; Bensaali, F.; Sotudeh, R., "OCR-based neural network for ANPR," in Imaging Systems and Techniques (IST), 2012 IEEE International Conference on , vol., no., pp.393-397, 16-17 July 2012.

[14] Shah, P.; Karamchandani, S.; Nadkar, T.; Gulechha, N.; Koli, K.; Lad, K., "OCR-based chassis-number recognition using artificial neural networks," in Vehicular

Electronics and Safety (ICVES), 2009 IEEE International Conference on , vol., no., pp.31-34, 11-12 Nov. 2009.

[15] Sahu, N.; Raman, N.K., "An efficient handwritten Devnagari character recognition system using neural network," in Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on , vol., no., pp.173-177, 22-23 March 2013.

[16] Wikipedia contributors. Histogram of oriented gradients. Wikipedia, The Free Encyclopedia. November 2, 2015, 18:21 UTC.

[17] Journal of the American Medical Informatics Association. (n.d.). Retrieved December 12, 2015, from http://jamia.oxfordjournals.org/content/21/5/871

[18] Christensson, P. (2006). *OCR Definition*. Retrieved 2015, Dec 11, from http://techterms.com

.