VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION

ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Feature Extraction and Feature Selection using Textual Analysis

Hemlata Badwal[1], Prof. Chandani Patel[2]
[1]*(Department of Computer Applications, VIVA School of MCA, Virar, Maharashtra, India)*
Email: badwalhemlata@gmail.com
[2]*(Department of Computer Applications, VIVA School of MCA, Virar, Maharashtra, India)*
Email: chandaniapatel@gmail.com

Abstract: *After pre-processing the images in character recognition systems, the images are segmented based on certain characteristics known as "features". The feature space identified for character recognition is however ranging across a huge dimensionality. To solve this problem of dimensionality, the feature selection and feature extraction methods are used. Hereby in this paper, we are going to discuss, the different techniques for feature extraction and feature selection and how these techniques are used to reduce the dimensionality of feature space to improve the performance of text categorization.*

Keywords *– Character Recognition, Feature Extraction, Feature Selection, Image Segmentation, Pre-processing.*

1. INTRODUCTION

The conversion of textual documents in digital form have increased rapidly worldwide. This is where the text classification becomes a dire need to handle these documents. The character recognition systems are thus used to fulfill these needs. The main objective of character recognition systems is to recognize and classify the text on the basis of predefined categories using some classifiers. Several analysis works have been done to evolve newer techniques and strategies that would scale back the time interval for processing whereas providing higher recognition accuracy.

In the following section, the four major steps used for handwriting recognition systems are described: - Pre-processing, Image segmentation, Feature extraction and Classification. Section II describes the various methods available for pre-processing, feature selection and feature extraction and the classifiers [1]. Section III presents the conclusive idea regarding the reviewed methods.

2. WORKING PRINCIPLE

Handwritten character recognition is the ability of an automated system to recognize handwritten character or sentence from photographs, documents, touchscreens and other such devices by a computer [1]. The image of the written sentence / word / character may be gathered either offline or on-line. In off line technique it may be from a scanned image of a paper. In on-line the sleuthing motion of the pen tip, for example by a pen-based display screen surface capturing temporal frequency [2].

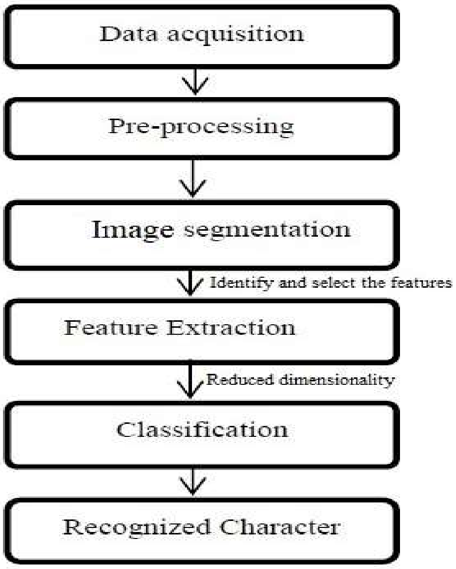

Fig. 1 Block Diagram of Character Recognition

Fig. 1 describes the process of text classification. Although, text classification has one limitation. This limitation includes high spatial property of feature area thanks to terribly sizable amount of options. The larger the options, the bigger the complexity of methods used for text classification and also the smaller the accuracy due to irrelevant or redundant terms within the feature space. Feature extraction and feature selection methods are used to eradicate this downside.

The initial major step in textual classification is Pre-processing. The moot data existing in the document is eliminated using the pre-processing. The subsequent step, image segmentation decomposes an image of a sequence of characters into sub images of individual symbols. Then, each character image is resized into m x n pixels towards the training network. These sub images are used to select the subset from the original feature set on the basis of importance of features, known as Feature Extraction. The fundamental task of feature extraction and selection is to identify a group of the most effective features for classification; so as to compress the high-dimensional feature space to low-dimensional feature space, for efficient classification [4] [16].

2.1 PRE-PROCESSING

When a document is scanned it requires some preliminary processing. This pre-processing helps to eliminate the irrelevant data in the scanned image in the form of noise and resize the image. For pre-processing the document, noise reduction, binarization, skew correction, etc. techniques can be used. The main objectives of pre-processing are [1]:

1. Noise reduction
2. Binarization
3. Skew correction
4. Stroke width normalization

Noise Reduction

The noise removal is achieved by converting RGB image into grayscale. It is then converted into binary image consisting only pixels 0's (white) and 1's (black). Unnecessary background pixels are removed from original image. To remove the redundant or irrelevant data, the threshold value is applied to the image.

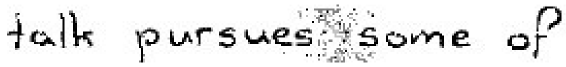
talk pursues some of

Fig. 2 Grayscale image with noise

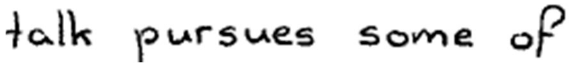
talk pursues some of

Fig. 3 Reduced noise image

Stroke-width normalization

After normalization generally it reduces the amount of data to be processed. For e.g. by thinning the shape information of a character can be gathered without losing the data.

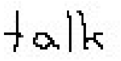

Fig. 4 Stroke-width normalization image

Skew Correction and Slant Removal

Skew correction methods are used for the alignment of the coordinate system of the scanner with respect to that of the document. Its main approaches include correlation, projection profiles and Hough transform etc [1]. The slant of any handwritten text(s) varies from one user to another. The characters can be normalized by using the slant removal methods [1].

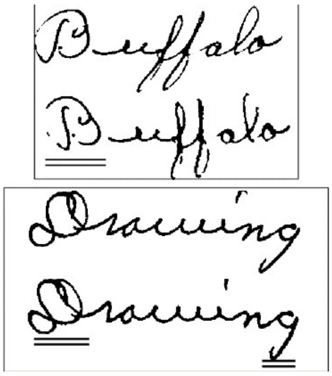

Fig. 5 Skew correction and Slant removal image

Binarization

In Binarization, a gray scale image if transformed to a binary image using global thresholding technique [3]. Let's assume f(x,y) is an input image. T is the threshold value and g(x,y) is the output image of thresholding process then the mathematical equation of this conversion is g(x,y) =1 if f(x,y) ≥ T otherwise 0.

Fig. 6 Binarized image

2.2 IMAGE SEGMENTATION

Image Segmentation involves two major steps [4]:

1. Line Detection
2. Word Detection

Line detection

Lines are differentiated using the Hough Transform, horizontal projections, etc. technique.

Hough transform detects straight lines. The straight line y = mx + b can be represented as a point (b, m) in the parameter space, where b represents the intercept parameter and m represents the slope parameter.

Horizontal Projection profile is the projection profile of an image along horizontal axis. It calculates each row for all column pixels in that row.

Fig. 7 Line Segmentation

Word detection

Words can be identified using vertical projections, connected component analysis, etc.

Vertical Projection profile is the projection profile of an image along vertical axis. Vertical projection profile is calculated for each column as summation of all row pixel values inside the column.

Connected component labeling works by scanning an image, pixel-by-pixel (from top to bottom and left to right) so as to identify connected pixel regions, [19] i.e. regions of adjacent pixels that share the same set of intensity values V. For a binary image V={1} [20].

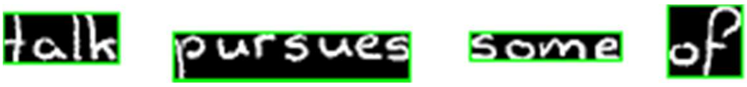

Fig. 8 Word Segmentation

2.3 FEATURE EXTRACTION

Feature extraction methods include principal component analysis (PCA), latent semantic indexing (LSI), etc [3].

Principal Component Analysis (PCA): Principal Component Analysis is dimension reduction technique that extracts the information from various dataset. PCA produces a set of lower-dimensional features from the original dataset. PCA is sensitive to the relative scaling of the original variables [5]. PCA determines the number of principal components using certain criterions. The dimensions representing the principal components at the best are chosen. The number of principal components to be chosen varies depending upon the quality of the dataset.

Latent Semantic Indexing (LSI): Latent Semantic Indexing (LSI) is a technique that projects queries and documents into a space with “latent” semantic dimensions. LSI has fewer dimensions than the original space. It works as a similarity matrix that is an alternative to word overlap measures like td.idf. LSI creates a matrix of words and sentences in a document that determine the occurrences of similar words which can be used for classification.

2.4 CLASSIFICATION

Classification is a supervised machine learning technique which identifies the category that a new observation belongs to, based on the observations used in the training set. The classifier utilization depends on several factors, such as available training set, range of free parameters. Some of the vital techniques which can be used for classification are- decision Tree (DT), k-Nearest Neighbor (k-NN), Bayes Classifier, Neural Networks (NN), Support Vector Machines (SVM), etc [1].

Decision Tree (DT) Classifier: In Decision tree classifier, the features are split into completely different regions corresponding to the classes available. It uses Information Gain value to identify the node at each level of the DT. The feature with the highest Information Gain is chosen as the node. The class is assigned along the path of the assigned nodes containing the features.

k-Nearest Neighbor (k-NN) Classifier: The purpose of k-NN classifier is used to assign the class label to a cluster of similar elements. It calculates the nearest neighbors using the Euclidean distance matrix. The similarity score is used as the weight of the classes of the neighbor document.

Naive Bayes (MB) Classifier: Naive Bayes Classifier is a probabilistic classifier. It assumes the independence of features. The value of a particular feature is independent of any other feature’s value, given the class variable [17]. For example, a fruit can be considered to be an apple if it is round, red, and about 10 cm in diameter. A naive Bayes classifier considers each and every feature contribute independently to the probability that this fruit is an apple, regardless of any attainable correlations between the color, roundness, and diameter features. Naive Bayes Classifier has surprisingly performed quite well in many complex real-world issues albeit it considers naive assumptions. Training and testing phase in Naive Bayes is easy for implementation and computation.

Support Vector Machine (SVM) Classifier: Support Vector Machine belongs to supervised learning classification technique. SVM algorithm consists of 2 types of versions: non-linear and linear versions [18]. In non-linear version, classes are not separated i.e. no straight lines that separate the classes can be found. In linear version, the classes are separated with the hyperplanes. Linear classifier is defined as wT x + b = 0. Where, w is the direction of the hyperplane and b is the precise position of hyperplane [18].

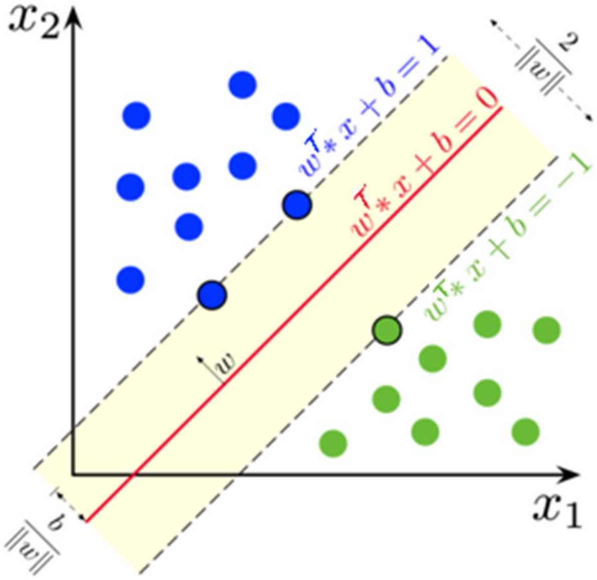

Fig. 9 Linear SVM Classifier

SVM find this hyperplane by using margins and support vectors. The region between the two hyperplanes wT x + b = 1 and wTx + b = -1 is known as margin. 2/||w|| is the width of the margin. The tuples that fall on the two hyperplanes are known as support vectors.

3. Conclusion

With the rapid increase in the digital conversion of texts, the need for handwriting recognition systems are soaring higher and higher. Here in this paper, we have discussed the techniques used in converting the text into the digital form. The accuracy rate of the systems determines how well a system is. But the text classification suffers from the higher dimensionality problem. To reduce the feature space and increase the accuracy of the systems, we have reviewed the pre-processing techniques, image segmentation, feature extraction as well as the classifiers available for text recognition.

References

[1] Vijay Prasad and Yumnam Jayanta Singh, "A study on method of feature extraction for Handwritten Character Recognition", *Indian Journal of Science and Technology*, pp. 174-178, *March 2013.*

[2] Ayush Purohit and Shardul Singh Chauhan, "A Literature Survey on Handwritten Character Recognition", *(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1), 2016.*

[3] J. Pradeep, E. Srinivasan and S. Himavathi, "Diagonal based feature extraction for Handwritten alphabets recognition system using neural network", *International Journal of Computer Science & Information Technology, Vol 3, No 1, Feb 2011.*

[4] Foram P. Shah and Vibha Patel, "A Review on Feature Selection and Feature Extraction for Text Classification", *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), March 2016.*

[5] Gaurav Kumar and Pradeep Kumar Bhatia, "A Detailed Review of Feature Extraction in Image Processing Systems", *Fourth International Conference on Advanced Computing & Communication Technologies, Feb 2014.*

[6] T. Fujisaki, T.E. Chefalas, J. Kim, C.C. Tappert and C.G. Wolf, "On-Line Run-On Character Recognizer: Design and Performance", *Character and Handwriting Recognition: Expanding Frontiers. P.S.P. Wang, ed.,* pp. 123-137, *Singapore: World Scientific, 1991.*

[7] Oivind Due Trier, Torfinn Taxt, Anil K. Jain, "Feature Extraction Methods for Character Recognition-A survey", *Pattern Recognition, Vol. 29,* pp. 641-662, *April 1996.*

[8] K. Gaurav and Bhatia P. K., "Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition", *2nd International Conference on Emerging Trends in Engineering & Management, ICETEM, 2013.*

[9] A. Brakensiek, J. Rottland, G. Rigoll, A. Kosmala, "Offline Handwriting Recognition using various Hybrid Modeling Techniques and Character N-Grams", Available at http://irs.ub.rug.nl/dbi/4357a84695495.

[10] R.G. Casey and E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No.7,* pp. 690-706, *July 1996.*

[11] S. N. Srihari, R. Plamondon, "On-line and off- line handwritten character recognition: A comprehensive survey," *IEEE. Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1,* pp. 63-84, *2000.*

[12] Harun Uguz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm", *Elsevier Knowledge-Based Systems,* pp. 1024-1032, *2012.*

[13] S. Niharika, V. Sneha Latha and D.R. Lavanya, "A Survey on Text Categorization", *International Journal of Computer Trends and Technology, Vol. 3,* pp. 39-45, *2012.*

[14] M. Blumenstein, H. Basli, B. Verma, "A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters", *Proceedings of the 7th International Conference on Document Analysis and Recognition, Vol. 1,* pp. 137–141, *2003.*

[15] Anshul Gupta, Manisha Srivastava, "Offline Handwritten Character Recognition", pp. 1-27, *April 2011.*

[16] Shifei Ding, Weikuan Jia, Chunyang Su, Fengxiang Jin, "A survey on Statistical Pattern Feature Extraction", *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, 4th International Conference on Intelligent Computing, ICIC 2008, Proceedings* (pp.701-708), *Shanghai, China, September 15-18, 2008.*

[17] Abdul Salam Shah, M.N.A. Khan, Fazli Subhan, Muhammad Fayaz, Asadullah Shah, "An Offline Signature Verification Technique Using Pixels Intensity Levels", *International Journal of Signal Processing, Image Processing and Pattern Recognition, August 2016.*

[18] Youness Tabii, Mohamed Lazaar, Mohammed Al Achhab, Nourddine Ennaya, Book872 ,Big Data ,Cloud and Applications :*Third International Conference, Communications in Computer and Information Science, Kenitra, Morocco, April 2018.*

[19] Available at https://homepages.inf.ed.ac.uk/rbf/HIPR2/label.htm

[20] Faiq Baji, Mihai L. Mocanu, Popa Didi Liliana, "Brain tumor detection based on asymmetry and K-means clustering MRI image segmentation", *Journal of Engineering Science and Technology, Vol. 13, No. 12*, pp. 4145 – 4159, *2018.*