International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue (4) July, Technovision-2014, ISSN 2249–071X

Technovision-2014: 1st International Conference at SITS, Narhe, Pune on April 5-6, 2014

Writer Identification using Binarized Features for Handwritten Marathi Numerals

Rini Mathew
Department of Electronics and Telecommunication,
Bharati Vidyapeeth's College of Engineering for Women,
Katraj University of Pune, Pune-411043, India
Email: rinimathew1604@gmail.com

Yugandhara Nimbalkar
Department of Electronics and Telecommunication,
Bharati Vidyapeeth's College of Engineering for Women,
Katraj University of Pune, Pune-411043, India
Email: yugandharanimbalkar@gmail.com

Tejaswini Patil
Department of Electronics and Telecommunication,
Bharati Vidyapeeth's College of Engineering for Women,
Katraj University of Pune, Pune-411043, India
Email: patil.tejaswini72@gmail.com

Prof. S. L. Kore
Guide, Associate Professor,
Department of Electronics and Telecommunication,
Bharati Vidyapeeth's College of Engineering for Women,
Katraj University of Pune, Pune-411043, India
Email: sharadakore@gmail.com

Abstract **– In this paper we show that we can extract a number of features from handwritten digits. These features are then used to identify or discriminate the writer. The features described are computational features which can be extracted by computer algorithms with a high degree of accuracy. These features cannot be computed by humans. All these features were appropriately binarized and binary feature vectors of constant lengths were formed. These vectors were then used for Hamming distance measure. The classification is done using Support Vector Machine (SVM). For this task a database of 50 writers was created. Each writer was asked to write Marathi numerals 0 to 9 five times. The results show that the combined features work well in discriminating the writers. Although this work is carried out for digits, it produces good results for isolated alphabets.**

Keywords **– Document Analysis, Feature Extraction, Writer Identification.**

I. Introduction

Writer identification has been an active research topic for several decades in the image processing and pattern recognition community [1], [2]. Writer identification is very essential these days due to its applications in different fields like historical document analysis [3]. Although the physiological biometric modalities like palm print, face, iris, fingerprint of a person is a strong identification technique, the identification of a person on the basis of handwriting samples still remains a useful biometric modality, mainly due to its applicability in the forensic field. In legal cases, document examiners are consulted for their opinions to prove the authenticity of the document. The authenticity of questioned document can be proved or disproved using various strategies [4]. This can be done by computing the quantitative as well as qualitative features. A lot of research has been carried out on handwriting based writer identification and also lot of problems and difficulties are encountered in this field. The aim of our work is to extract numerous features which can be computed by computer algorithms only. We try to find information from a person's handwriting which will make it identifiable from other person's handwriting and distinguishable from others. Our work proves useful in commercial applications like reading bank cheques to find the writer of the cheque. It can also be used for historical document analysis.

II. Overview of Existing Methods

Various works have been carried out in previous years on writer identification. Said, Tan and Baker proposed the texture analysis approach to writer identification wherein they took the handwriting as an image containing some special texture, and applied the well-established 2-D Gabor filtering technique to extract features of such textures [5]. Long Zuo, Yunhong Wang, Tieniu Tan applied Principal Component Analysis to the gray-scale handwriting images to find a set of individual words which best characterize a person's handwriting style and have maximal difference from other people style[6]. Shrihari and Cha [7] extracted twelve shape features from the handwriting text lines to represent personal handwriting style. The features mainly contained visible characteristics of the handwriting, such as width, slant and height of the main writing zones. These are the method that has been applied to a line of text. In this paper we are working on isolated digits. Not much work has been carried out specifically for finding out discrimination data for handwritten digits.

III. Methodology

The scanned images of handwritten digits are preprocessed i.e. noise removal from the images take place and they are appropriately binarized and features described in (B) are extracted and classified to discriminate the writer.

A. Database:

A database of 50 writers is created for this purpose.

Each writer was asked to write Marathi numerals 0 to 9 five times. The handwritten numerals were then scanned at 300 dpi. The feature extraction algorithms were applied to all the handwritten samples and results were obtained accordingly.

B. Computational features:

i. Aspect ratio –

This is simply calculated as the ratio of height and weight. See Fig. 2.

ii. Zero Crossings

Zero crossing is nothing but the number of transitions from white to black or from black to white in a digit image. For finding out this feature, the digit height is divided into number of blocks (say 5 or 8) and then the zero crossing value is calculated along each of the 5 or 8 lines either in the horizontal or in the vertical direction at a time.

iii. Number of Loops

1. Degree of Roundness – Roundness is the measure of how closely the shape of an object approaches that of a circle. For a perfect circle the degree of roundness is zero.

2. Area of a loop- It is defined as the number of white pixels contained inside it.

3. Loop Length- A length of line that is curved or doubled over making an opening is a loop. The loop length is found by first detecting the contours and then calculating the number of black pixels left behind in the image.

4. Slant – It gives the angle at which the loop is inclined. It is calculated by dividing a loop into two parts with respect to its height and then centers of gravity of the two halves are found. The line connecting these gravity centers is considered to represent the loop angle.

5. Loop Fissure Length –When the loop is incomplete or open then this feature is used. This feature gives the total number of missing pixels in a loop. The feature value is normalized by dividing the number of missing pixels by the loop length calculated in the feature of loop length. The feature value becomes independent of the character or digit size due to normalization.

iv. Slant

This feature gives approximate slant of the numeral. For example, in the case of the digit "4" slant can be taken as the tangent of an angle made by the east most point and the end point. Similarly, we can develop rules for other digits and characters.

v. Fixed Point distance and angular measure –

For this feature, image is first resized in a standard size. In our work it is divided into 6x4. Then fixed point was chosen to be the first pixel (mostly white) of the bounding box of the digit. From this pixel the distance of each black pixel was calculated.

vi. Number of Junctions –

This feature describes the number of junctions present in the numeral. A junction is a point (black pixel) having three 8-connected neighbours (minimum) that are black. It is formed if two lines intersect each other. This operation is to be performed on a thinned image. In fig. 4 the numeral shown has only one junction.

vii. Number of End-Points –

This feature is described by the number of pointed ends present in the numeral. It is found by checking all the black pixels that do not have more than one 8-connected neighbour. In Fig. 2 there are 3 end points in below numeral.

viii. Width / Height Distribution

Width/Height distribution is the change in the width or height of the character as a digit image is traversed across in horizontal and vertical direction along the lines in feature of zero crossing.

ix. Pixel Density

In this feature, firstly, bounding box of the digit is divided into 24 equal rectangles of equal area. Then the number of black pixels in each box is then found and divided by the rectangle area. See fig. 1.

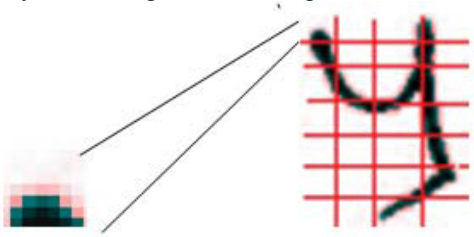

Fig.1. Pixel Density

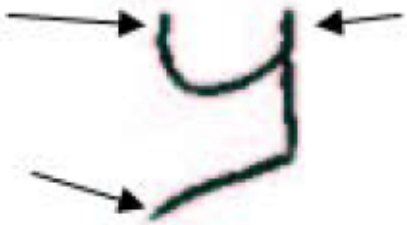

Fig. 2: No. of end-points

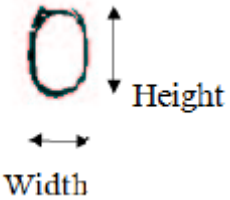

Fig.3. Aspect ratio

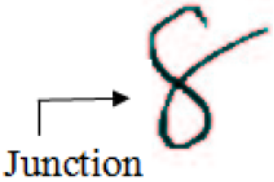

Fig.4. No. of junctions

x. Centre of Gravity

In this feature, centre of gravity of the bounding box of a digit is calculated. Then the ratio is taken of the X-coordinates of the centre of gravity with the Height and Y-coordinates of the centre of gravity with the width of the bounding box respectively are stored as a feature.

xi. Gradient Features

Similar to feature no. v and ix, the digit image is divided into 6x4 blocks. Then the gradients of each black pixel of the original image are calculated using a simple 3x3 Sobel operator. The gradient measures the magnitude and direction of the greatest change in intensity in a small neighborhood of each pixel. This feature can also provide slant of the character.

All the features from i to xi were extracted and appropriately binarized. Each of the features mentioned were binarized taking into consideration their nature. Once binary feature vectors for all the digits of each writer are formed, the data can be used for the purpose of author discrimination by means of various classification techniques. In our work we have used the Support Vector Machine (SVM) Classifier.

IV. Experimental Results

Table 1: Feature values

Feature number	Corresponding no. of values obtained
i	1
ii	40
iii	1
iv	1
v	23
vi	1
vii	1
viii	24
ix	2
x	24
xi	118

We extracted the features mentioned above and 118 values for the features indicated in the table were obtained. Since each writer has written each digit 5 times we have 5 feature vectors per digit per writer. The features were then binarized. This set was then divided into standard and test patterns. The Hamming distance metric for feature vector distance computation was used. For any given test vector its distance was calculated with each vector of each writer in the standard vector set. The digit test vector was finally assigned to the writer which produced the least average Hamming distance. If there existed two such sets for two different digits having the same average Hamming distance, the test vector was rejected by the algorithm. In this way accuracy was calculated.

V. Conclusion and Future Scope

A set of features were extracted from a form of handwritten digits. A feature set was formed which can be used to prove the authorship of the handwritten documents containing digits. All the features are computational features. They were extracted using computer algorithms with high accuracy. The accuracy obtained is high because the features which are being extracted characterize the handwriting of a person. In other words they are the innovative characteristics that prove the identity of a person. Also the classification done due to SVM Classifiers makes our system efficient. Although the work was carried out for handwritten digits it can also be implemented for isolated characters.

Acknowledgment

We express our sincere thanks to our project guide Prof. S. L. KORE who always being with presence & constant, constructive criticism to made project successful. I would also like to thank all the staff of Electronics and Telecommunication Department for their valuable guidance, suggestion and support through the project work, who has given co-operation for the project with personal attention. We again take it as great privilege to express our heartfelt thanks to our principal Dr. D.S. BILGI and Head of Department Prof. S.T. KHOT for their valuable suggestion for developing project at every state. Above all we express our deepest gratitude to all of them for their kind-hearted support which helped us a lot during project development. They offered us plenty of opportunities while working with them, rendered us in valuable help & helped us linking practical knowledge with theoretical one taught to us in our college.

References

[1] R. Plamondon and G. Lorette, “Automatic signature verification and writer identification – the state of the art,” Pattern Recognition, vol. 22, no. 2, pp. 107–131, 1989.

[2] A. Schlapbach, Writer Identification and verification. Dissertations in Artificial Intelligence, 2008, vol. 311. 148 pages.

[3] Sreeraj. M, Sumam and Mary Idicula “A Survey on Writer Identification Schemes” International Journal of Computer Applications (0975 – 8887) Volume 26– No.2, July 2011

[4] Hilton, O., “Scientific Examination of Questioned Documents – Revised Edition”, CRC Press, Inc., 1993.

[5] H. E. S. Said, T. N. Tan and K. D. Baker, “Personal Identification Based on Handwriting”, Pattern Recognition, vol.33, no.1, pp.149-160, 2000.

[6] Long Zuo, Yunhong Wang, Tieniu Tan ‘Personal Handwriting Identification Based on PCA’ National Laboratory of Pattern Recognition (NLPR) Institute of Automation, Chinese Academy of Sciences. S. Cha and S. N. Srihari, “Multiple Features.

[7] Integration for Writer Verification”, the Proceedings of 7th IWFHR2000, Amsterdam, Netherlands, Sept. 2000, p 333-342.

[8] Graham Leedham and Sumit Chachra “Writer Identification using Innovative Binarised Features of Handwritten Numerals” Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR’03).