iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** XII **Month of publication:** December 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

An Intelligent System For Writer Identification Of Devanagari Handwritten Documents Using ANN

Praveen Sagar[1], Akhilesh Pandey[2]
[1]M.Tech. Scholar, [2]Associate Professor, Department of Computer Science & Information Technology
SGVU, Jaipur, India

Abstract— Many studies about writer identification have been done on different handwritten documents but less no of researches has been reported on Hindi or Devanagari documents. Here, in this technique, we accept handwriting as texture image and feature extraction done from preprocessed documents image. Significantly, the feature of presented method uses feature extraction which is suitable for structure of Hindi handwritten texts. And a new feature extraction technique is also presented which is based on classifier design. Firstly, we study different methods for feature extraction from output of preprocessing. This paper represents performance of slant estimation techniques on Devanagari handwritten documents. Slant is a unique feature; it acts as a vital role in writer identification, OCR and graphology systems. For estimation of slant, Radon Transform and Hough Transforms are applied on the word images. Further to improve the performance of these techniques, the extracted images are reconstructed using Zernike moments. For conducting this experiment, samples available on CPAR-2012 for Hindi datasets are used. The experimental result shows that slant estimated from images reconstructed using Zernike moment gives better results.

Keywords—Slant, Writer identification, Radon transform, Hough transform, Zernike moments, Handwriting analysis.

I. INTRODUCTION

Handwriting is the essential measure for communication which is also a measure for identification of any person who can write, it helps to identify the writer on the basis of their writing style, font, and other aspects. These unique appearances of handwriting separate writer's writing from another's; even it may be possible to recognize the author, if the same copy book is shared by two writings. This automatic writer identification possesses a valued resolution for the forensic experts, document examiners and graphologists to writer recognition.

A. Writer Identification As A Biometric

The handwriting is being acknowledged as a legal biometric, for, the evidence of Forensic Document Examiners or Questioned Document Examiners (FDEs/QDEs) which has been judged with authorization allowable. Expert witness mainly use these facts regarding the genuineness of specific writings, and for defining the source of a historical writing where this is unidentified or in demand.

B. Writer Identification Process

The matching process of writing features with available samples written by authors known as writer identification.

1) Pre-Processing: To work with this phase as an appropriate gathering and ground work of data, and of course very detailed to the type of data is being classified. This phase includes the assortment and training of the data into an appropriate practice to work with. In terms of images, this step scanning, digitizing, manual or automatic cleaning and enrichment can be included.
2) Feature Extraction: There is naturally far too much evidence into the input data to categorize it directly, and much of this information is inappropriate to defining the correct class. Therefore the pattern recognition procedure depends on determining or extracting components of the actual data which are useful indicators of the class from it is actually belongs. These components are well-known as features, where its expansion forms a key measure of every pattern recognition system area. In the field of writer identification, the main part of the research has been concentrated in this phase. In addition, there are several optional phases which may be accepted to deriving enhanced features from those primarily mined.
3) Comparison and Identification: On the basis of feature extraction data to be classified in different set of classes which contain similar type of data. After that each data to be compared with available samples and the maximum number or percentage of

similarities indicates the writer of particular handwritten document.

II. LITRATURE REVIEW

Various research work has been suggested beginning like a series from a new feature collection [1], discovering unlike feature grouping systems [2] and examining the compassion of typescripts for writer identification schemes [3]. These research works are based on slant involved features for writer identification system. Handwritings having salient property known as Slant. The style of pen hold and the relative assistances of wrist with finger actions are the cause of Slant. Writer verification uses Slant which appears to be a crucial feature: it acts a significant role in biometric schemes, as it is a key component of angular features [6] for recognition of individual writer. It is not unforeseen that slant adjustment is the furthermost commonly applied kind of disguise [7], meanwhile humans can effortlessly adjust the slant in the course of writing, and the outcome on the visual presence is affected [8]. Consequently, a significant step for construction biometric systems tough for disguise is by modifying the slant. A noticeable methodology is to achieve the correction by make over the image by way of the shear process, possibly bring about in the writer's regular handwriting. The motivation of this revision is twofold. First of all motivation is to decide how much statistics about the writer's distinctiveness is enclosed in the slant feature of normal writing. In the first investigation slant testing will be verified by reducing the slant in normal handwriting (slant dismissal) and determining to what level the performance of programmed writer verification damages. This research gives to the hypothetical foundation for computational writer features grounded for directionality, such as the Turning point (Hinge) feature [9]. It needs an uncharacteristic movement of the finger and wrist coordination, which may disturb the curvature. It has been recommended that there must be likewise an influence on writing speediness, joining strokes, pressure, flair, structure, and size [10]. Moreover, masked writing is less reliable [10]. At the second motivation, to what level of such additional effects arise? It may be quantitatively resolute. This may be completed by clipping slanted script back to the made up writer's normal slant perspective (slant improvement) with defining the presentation of writer verification by means of state of the art structures. This is the first stage in planning a new biometric organizations which is tough to masquerade.

Writer identification in monolingual handwritten writings has been well thought out together with Chinese [11], Telugu [13], Japanese [12], latin [14-17], Oriya [18], Bengali [17], and Arabic [19]. The expansion of Devnagari handwriting based scheme is a comparatively recent sensation. On the other hand, there have been certain hard work done in this trend. Mukherji et al. [22] have defined the improvement of a handwriting confirmation system aimed for Devnagari handwriting and Vighnesh et al. [23] suggested a system for person identification with the help of Hindi typescripts and words. Both of them have used predictable pattern acknowledgement procedures. In their handwriting recognition systems, with the help of some file level with word level structures like inter-word spacing, height and width of script strokes, word slant, and average word orientation. Many of work previously tested their method on Devnagari handwriting, which was aggregate of 50 pages that composed from 10 different writers, and the final trained and confirmed their scheme on 18,000 words of amounts changeable from 2–4 letters handwritten by 50 authors.Both of these systems reached nearby to 90% identification accurateness, on the other hand their conservative strategies lack feature detection support. Handwriting specialists bring up to a comprehensive list of structural assets as a standard feature dataset for discovering the vagueness of handwriting. To the best of our understanding, no related experimentation has been implemented in the past. The research will be executed on a once more generated public dataset: CPAR-2012 dataset, comprising both ordinary and slanted handwritten scripts of a number of matters.

III. PROBLEM IDENTIFICATION

A. Data Collaction

There is a small amount of dataset available for Hindi manuscript. The Hindi language recognition compare to latin language recognition can be a cause for very limited progress. To overcome the shortcomings, a standard dataset CPAR-20l2 dataset of Devnagari script to be used for standard studies. The CPAR-20l2 dataset is the collection of images of constrained, semi-constrained and unconstrained handwritten digits; isolated fonts; unconstrained and constrained pangram script; digitized data assembly arrangements. We have completely settled a pangram of Devnagari script to improve and validate practices for handwritten script analysis applications, like writer identification. The pangram dataset is standard which contains thirteen most commonly used vowels, fourteen changers and thirty six consonants. The pangram dataset is a collection of writer facts alongside the unrestricted and restricted handwritten pangram pictures. To evaluate the strength of writer identification practices, three samples have been collected from the same writer.

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

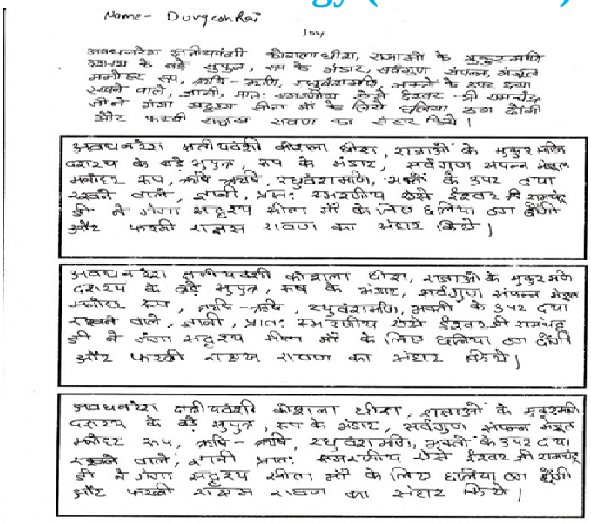


Figure 1. Handwritten text dataset for Hindi

B. Pre-Processing

Some noises like salt & pepper, Gaussian noise alongside scan lines are detected during data collection. Pre-processing of the image is much compulsory for eliminating these kinds of noises. For separation of the foreground and background, the pictures are binarized which helps in detection. Bounding Box of pictures are used to segmenting these images, which provides us the size of the smallest rectangle enclosing a region as well as co-ordinates. In this situation, exclusively coupled components are segmented through Bounding Box, which gives segmented components by extracting, cropping and storing.

C. Words And Letters Filtering

The image segments are separated into letters, words and other punctuation symbols after pre-processing. For two thresholds and are well defined based on manual clarifications for every one data model.

For I = 1 to n

{

If width (I_p) > th_1

$A_w(i) = I_p$

else if width (I_p) < th_1 && width (I_p) > th_2

$A_l(i) = I_p$

else if width (I_p) < th_2

$A_o(i) = I_p$

}

Where, n is the total number of components, is the pre-processed image, is a cell containing words, is a cell containing letters and is a cell containing components like dots and commas. Further is not considered because the images it contains are too small to have any impact on the slant angles of the document.

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

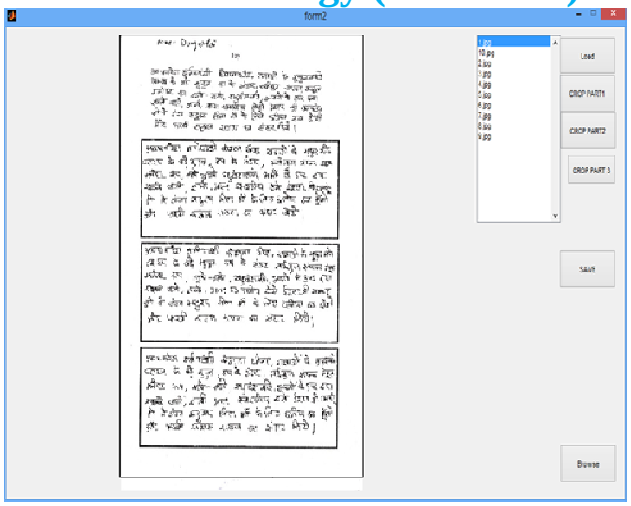


Figure 2. Graphical User interface design for dataset development.

A Graphical User Interface shown (Fig.2) here which was aimed to cover abstract data from the original image at present there was no current program that had these type of availability to practice in a simplified fashion. The interface will deliver capability to upload several types of images into the MATLAB background. Subsequently uploading the image with low pass filters were used to routine the loaded data. After dealing out with these arrangements a matfile will be produced and this set aside for advance processing. After that input document image is cropped and stored in the form of relational database as shown in fig. 3.

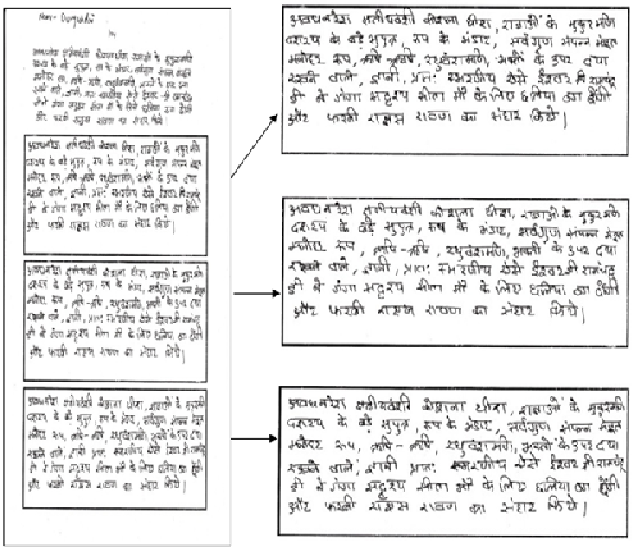


Figure 3. Original image is cropped and handwritten text is cropped and organized in relational database model.

IV. IMPLEMENTATION

The Proposed methodology is done in following five phases:
Input
Preprocessing
Word Slant Estimation
Classification
Writer Identification

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A. Input

A scanned copy of handwritten document to be stored as input for further pre-processing.

B. Preprocessing

Input image contains various types of noises which may occurred during the data acquisition process. These noises are removed by using Gaussian fi1ter. A 2-D, an isotropic (i.e. circularly symmetric) Gaussian has the form:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad \ldots (1)$$

These preprocessed images are binarized for further processing.

1) *Data Segmentation And Cleaning:* Binarise original images directly to an appropriate threshold in this phase can be found where applicable, or so they too closely composed in greyscale map, or the use of colour, because the image is too noisy, or non - text element are included.
2) *Finding Connected-Components:* A connected-component is a comprehensive piece of joined ink trace, delineated only by somewhere the author has elevated the pen. Extracting graphemes to be done by occasionally united with text-line extraction.

C. Word Slant Estimation

1) *Slant Estimation Using Radon Transform:* The radon transform computes projection of an image from a specific direction taken from different angles by rotating the source of projection around the centre of the image. The projection of a two dimensional image f(x, y) is computed at an angle θ with the equation:

$$R_\theta(x') = \int_{-\infty}^{\infty} f(x'\cos\theta - y'\sin\theta, x'\sin\theta - y'\cos\theta)dy \quad \ldots (2)$$

Where θ is the angle of rotation and

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad \ldots (3)$$

This transform detects a line at an angle. Using this transform, we computed the maximum value of Radon Transform over all lines forming angles $-60 \leq \theta \leq 60$.

2) *Slant Estimation From Hough Transform:* Hough transform is a feature extraction technique used in image analysis, computer vision and digital image processing. The purpose of the technique is to find imperfect instances of objects within a certain class of shapes by a voting procedure. This voting procedure is carried out in a parameter space. It is a method of estimating the parameters of a shape from its boundary points and is one global powerful method for detecting edges (lines and parametric curves). In general, the straight line y = mx + b can be represented as a point (b, m) in the parameter space. However, vertical lines pose a problem. They would give rise to unbounded values of the slope parameter m. Thus, for computational reasons, Duda and Hart proposed the use of the Hesse normal form

$$r = x\cos\theta + y\sin\theta \quad \ldots (4)$$

Where r is the distance from the origin to the closest point on the straight line, and θ the angle between the axis and that closest point.

3) *Zernike Moments:* The Zernike moments are a set of polynomials defined on a unit circle. These are calculated by using a set of Polynomials that form a complete orthogonal base defined in the unit circle $(x^2+y^2)\leq 1$. These moments are projections of the input image, f(x, y) in the space of the orthogonal functions. In general, moments describe numeric quantities at some distance from reference point or axis. For a discrete image, the Zernike moments of order n and repetition m are given by:

$$A_{n,m} = \frac{n+1}{\pi} \cdot \sum_x \sum_y f(x,y) \cdot V_{n,m}(x,y)^* \quad \ldots (5)$$

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Where, $(x^2+y^2) \leq 1$, $n \geq 0$, $|m| \leq n$, $n-|m|$ is odd and the symbol * represents the conjugated complex operator. The magnitudes $|A_{(n,m)}|$ are invariable to this rotation (this is, $A_{(n,m)} = A_{(n-m)}$).

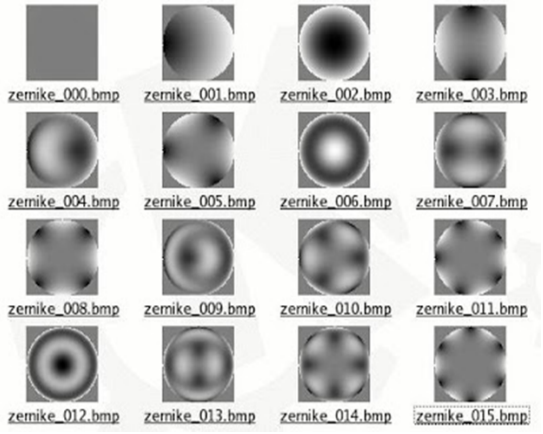


Figure 4. Zernike moments of various levels.

D. Classification

After understanding the features or attributes of a class are acknowledged, distinct objects might be well-known as to be appropriate or not to be appropriate to that class. The objects are allotted to classes by detecting patterns of differentiating characteristics and relating with related model participant to each class. Pattern recognition involves the extraction of patterns from data, their analysis and, finally, the identification of the category (class) each of the pattern belongs to. A typical pattern recognition system contains a sensor, a pre-processing mechanism (segmentation), a feature extraction mechanism (manual or automated), and a classification or description algorithm. In this project following classifiers are used.

1) KNN Classifier: The KNN classifier foretells the class sticky label of the test design x as of predefined class. The classifier catches the k closest neighbor of x and finds the class label of x using majority voting. The performance of KNN classifier depends on the choice of k and distance metric used to measure the neighbor distances. In this experiment used Euclidean distance metric (as shown in Eq.(6)

$$dist(X_i, Y_j) = \sqrt{\sum_{i=a}^{k} (X_{ja} - X_{ia})^2} \quad \ldots (6)$$

2) Neural Network Classifier: An N-layered feed-forward multilayer neural network contains one input (the first) layer, one output (the last) layer, and N- hidden (intermediate) layers. In this experiment neural network were created with 20 hidden layers in all the neural network classifiers. This classifier uses logsig transfer function. This functions calculating the layers output from its input. The output layer of feed forward neural network is given by Eq. (6).

$$O_i^k = f(net_i^k) = \text{logsig}(net_i^k) = \frac{1}{1 + e^{-net_i^k}} \ldots (7)$$

E. Writer Identification

After features extraction features are classified using the classification techniques. The accuracy of this methodologies can be seen by using confusion matrix, which is generated after comparison of trained datasets with testing datasets. The maximum number or percentage of similarities of features indicates the writer of a particular handwritten document.

V. RESULTS

These samples were composed by 10 different writers as three samples for each writer. The input image is segmented and binarised. For calculating slant, Radon transform is used. Following results came at time of following phases:

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A. Pre-Processing

1) Segmentation: The above image is segmented part of original image for binarization.

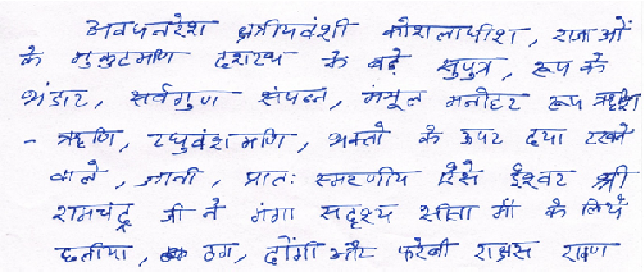


Figure 5. Cropped Input Image

2) Binarization: This image is in binarized form for removal of noises which is arises at the time of data acquisition process. This helps to filter words and letters.

3) Words And Letter Filtering: The binarised images is further segmented for filtering words, letters and punctuation marks.

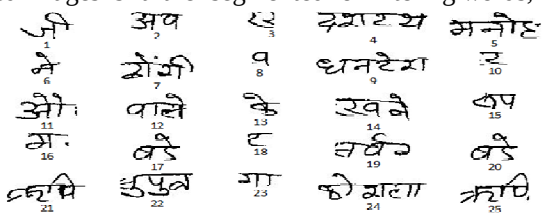


Figure 6. Cropped words and letters

B. Slant Estimation

After that Radon Transform is applied on the objects to obtain slant estimation of handwritten documents. The next figure shows the slant variation of three different writers.

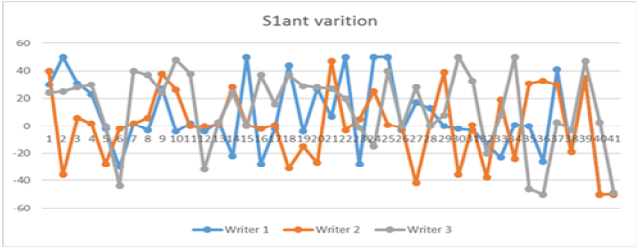


Figure 7. Slant based feature for three different writer

C. Classification

After slant estimation the objects to be classified to belonging classes. The confusion matrix to be generated which contains comparison statics of trained dataset (Output class) and testing dataset (Target class).

D. Writer Identification

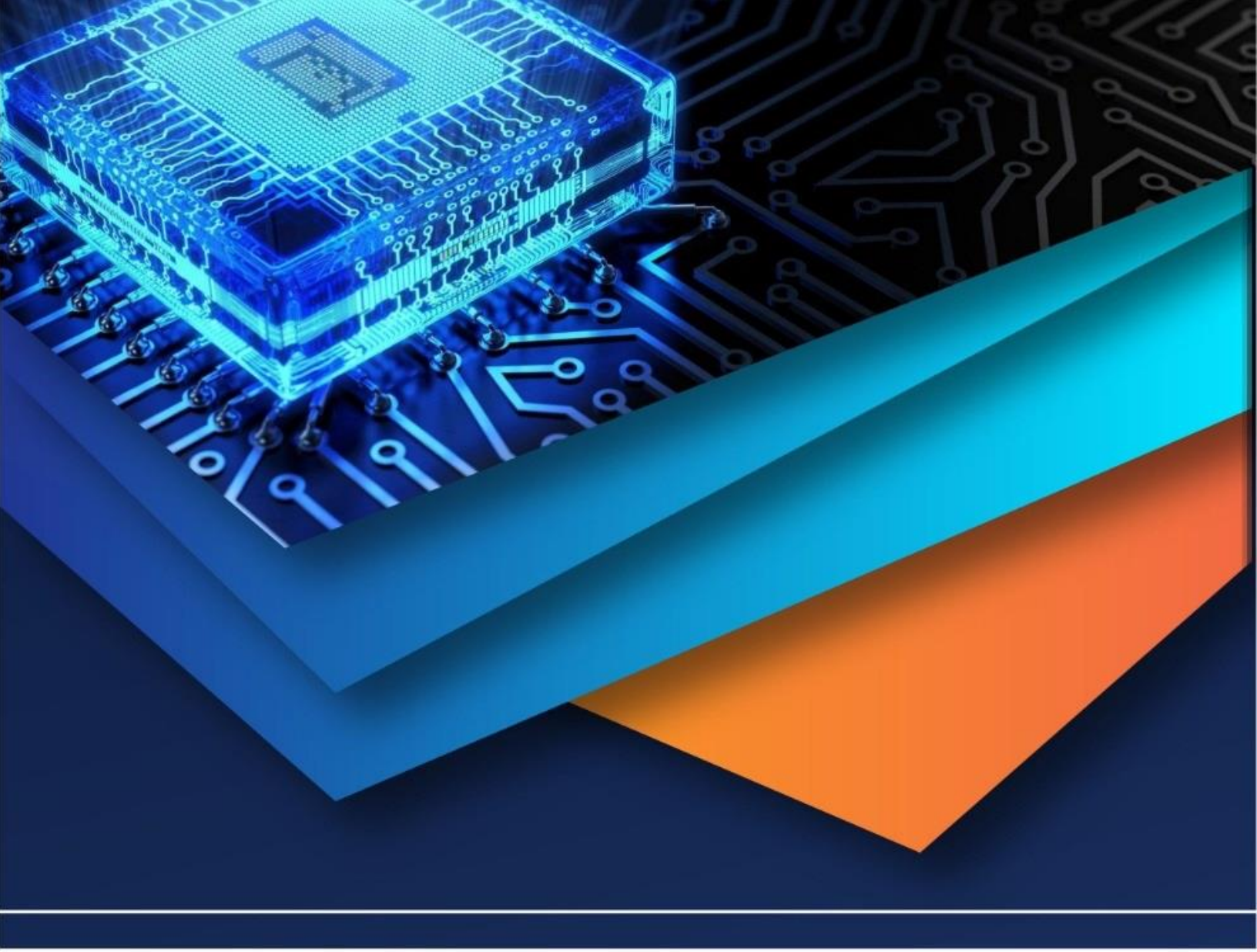
The maximum number or percentage of similarities indicates the writer of particular handwritten document. Based on above results it is observed that slant based writer identification in Hindi text. Comparing feature selection (slant) methods against each other, there are again many similarities across the experiments. Extremes of similarity range Hough based slant selection perform poorly across all experiments. Radon transform based slant estimation performs better in all experiments on word level. This supports the hypothesis that radon transform based slant estimation can be used for writer identification system. However, this distinction is not significant on the large-scale medieval-dataset experiment.

VI. CONCLUSION

This proposal is an exertion which presenting a writer identification system constructed on slant assessment with help of Radon transform. In this research it is clearly specified that transformation of slant gives good outcome whereas it is verified on unchanged qualified dataset although on dissimilar test sets its outcome differs. This obviously indicates that the working out set have impact on recognition accurateness. Additionally these slant assessment system will assist us in sentiment recognition, and skew adjustment. The outcome achieved in this proposal also illustrate that whereas it is skilled and practiced with KNN classifier on equivalent dataset provides l00 % accurateness. But in the case of other datasets when it is tried, it provides approx. 70% accurateness. The accurateness can be more enhanced by means of Fuzzy based methodology. On the basis of these facts it can be supposed that slant is a striking feature of handwritten writings. But it is not much explanatory for handwriting evaluation as is commonly supposed at the same time. This was seen in a sequence of writer verification research via three state-of-the-art statistical features: Baseline, Directions, Average Upper mid relation etc. can be collective combined to develop more precise outcomes.

REFERENCES

[1] Brink, A. A., Niels, R. M. J., van Batenburg, R. A., van Den Heuvel, C. E., & Schomaker, l. R. B, Towards robust writer verification by correcting unnatural slant, Pattern Recognition letters, 20ll, 32(3), 449-457.
[2] linne R. Mooney. Chaucer's Scribe. Speculum, 8l(0l):97–l38, 2006.
[3] Tom Davis. The Practice of Handwriting Identification. library, 8(3):25l–276, 2007. doi: l0.l093/library/8.3.25l.
[4] A. Jain. On-line signature verification. Pattern Recognition, 35(l2):2963–2972, 2002. doi: l0.l0l6/S003l-3203(0l)00240-0.
[5] L. Schomaker, Advances in Writer identification and verification, in: Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR), Sept. 23-26, 2007, pp. l268 – l273.
[6] M. Bulacu and l. Schomaker, Writer Style from Oriented Edge Fragments, in: Proceedings of the l0th Int'l Conf. Computer Analysis of Images and Patterns, 2003, pp. 460-469.
[7] Nickell, J., 2007. Detecting Forgery: Forensic Investigation of Documents. University Press of Kentucky.
[8] Koppenhaver, K.M., 2007. Forensic Document Examination: Principles and Practice. Humana Press. www.springerlink.com/content/r548gl8095452072/
[9] M. Bulacu and l. Schomaker, Writer Style from Oriented Edge Fragments, in: Proceedings of the l0th Int'l Conf. Computer Analysis of Images and Patterns, 2003, pp. 460-469.
[10] Forensic Handwriting Identification. Academic Press.
[11] l. Schomaker, M. Bulacu, and K. Franke, Automatic Writer Identification Using Fragmented Connected-Component Contours, in: Proceedings of the Ninth Int'l Workshop Frontiers in Handwriting Recognition (IWFHR), 2004, pp. l85-l90.
[12] C. Tomai, B. Zhang and S. Srihari, Discriminatory Power of Handwritten Words for Writer Recognition, in: Proceedings of the l7th Int'l Conf. Pattern Recognition, 2004, pp. 638-64l.
[13] A. Schlapbach, H. Bunke, Using HMM based recognizers for writer identification and verification, in: Proceedings of the International Workshop on Frontiers in Handwriting Recognition, IWFHR, Tokyo, 2004, pp. l67–l72.
[14] M. Bulacu, l. Schomaker, A Comparison of Clustering Methods for Writer Identification and Verification, in: Proceedings of the Eighth Int'l Conf. Document Analysis and Recognition, vol. 2, 2005, pp. l275-l279.
[15] A. Bensefia, T. Paquet, l. Heutte, A Writer Identification and Verification System, Pattern Recognition letters, 26(l0) (2005) 2080-2092.
[16] A. Bensefia, T. Paquet, l. Heutte, Handwritten Document Analysis for Automatic Writer Recognition, Electronic letters on Computer Vision and Image Analysis, 5(2) 2005 72-86.
[17] A. Schlapbach, l. Marcus and H. Bunke, A writer identification system for on-line whiteboard data, Pattern Recognition Journal 4l (7) (2008) 2382l–23897.
[18] S. N. Srihari, S. H. Cha and S. lee, Establishing Handwriting Individuality Using Pattern Recognition Techniques, in Proceedings of the Sixth International Conference on Document Analysis and Recognition, 200l, pp. ll95-l204.
[19] M. Bulacu and l. Schomaker, Writer Style from Oriented Edge Fragments, in: Proceedings of the l0th Int'l Conf. Computer Analysis of Images and Patterns, 2003, pp. 460-469.
[20] R. Niels, F. Gootjen and l. Vuurpijl, Writer Identification through Information Retrieval: The Allograph Weight Vector, International Conference on Frontiers in Handwriting Recognition, 2008, pp. 48l-486.
[21] M. Sreeraj and Sumam Mary Idicula, A Survey on Writer Identification Schemes, International Journal of Computer Applications, 26(2) (20ll) 23-33.
[22] P. Mukherji, P. P. Rege, and l.K. Pradhan, Analytical Handwriting Verification System for Devnagari Script, in: Proceedings of the Sixth IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP 2006), Palma De Mallorca, Spain., August 28 – 30, 2006.
[23] B. Aditya Vighnesh, B. Manikanta Yadav, Abhishek Kumar, M.V. Raghunadh, Autonomous Bilingual Character Recognition and Writer Identification, International Journal of Engineering and Innovative Technology, 2(l0) (20l3) 2l9-224.
[24] S. R. Deans, The Radon transform and some of its applications, John Wiley &amp Sons, l983.
[25] R. Kumar, A. Kumar, P. Ahmed, A Benchmark Dataset for Devnagari Document Recognition Research, in: Proceedings Int'l Conf. on Recent Advances in Telecommunications, Signals and Systems, 20l3, pp. 258-263.
[26] Pandey, A., Kumar, A., Kumar, R., & Tiwari, A. Handwritten Devanagri Number Recognition using Majority Voting Scheme. International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN, 2249-9555.
[27]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY