iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 5 **Issue:** XI **Month of publication:** November 2017

**DOI:**

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887*
*Volume 5 Issue XI November 2017- Available at www.ijraset.com*

# Unconstrained Character Classification For Off-Line Handwritten Devanagari Script Using Multi-Class Model For SVM

Deepu Kumar[1], Divya Gupta[2]
[1]*M.Tech Scholar,* [2]*Asst. Professor (Computer Science & Engineering),Ajay Kumar Garg Engineering College,Ghaziabad, India,*

***Abstract:*** ***Hindi script is being used in various languages, for instance Marathi, Rajasthan, Sanskrit, and Nepali and it is also the script of Devanagari. It is observed that errors in classification mainly due to complex structures, incorrect segmentation and high unevenness in writing styles, classification of characters from the unconstrained script has become a burning vicinity of interest for researchers. Computer-based pattern recognition is a process that involves preprocessing, feature extraction, feature selection, and classification. In that article, we have extracted features from HOG, the novelty of this approach to attain better accuracy and reduce misclassification as well as for classification of handwritten characters with a multiclass model for SVM. Implementation has been performed using a self-created dataset of 40 users for handwritten Hindi characters. The experimental results obtained from this self-created dataset described the effectiveness of this system. The proposed system has faster speed and higher accuracy than the traditional Hindi OCR's. Enormous applications and the future necessities of optical character recognition area open new paths for researchers. An effort is made to address the most crucial consequences and it is also tried to foreground the better directions of the research till date. Our experimental results present the high performance of these features when classified using SVM classification.***

***Keywords: Handwritten Devanagari Character, OCR, Feature Extraction, Classification, Classification Accuracy Cross-validation, Confusion Matrix.***

## I. INTRODUCTION

Off-line recognition of handwritten devanagari characters is a procedure of automatic computer recognition of optically scanned characters. Several OCR's are available commercially in the market [1][3]. It can thus immensely lead to the advancement of automation processes and can improve the interface between human and machine in various applications. Off-line recognition of hand written devanagari character is one of the crucial and challenging areas of pattern recognition and more specifically in document image analysis. Some practical applications of devanagari character recognition systems are: (1) processing cheese without human involvement, (2) reading aid for the blind, (3) automatic text entry into the computer for desktop publication, library cataloguing, ledger, (4) automatic interpretation of city names and addresses for postal mail, (5) document data compression etc. There is a great need for OCR related research in Indian scripts, even though there are many technical challenges as well as the lack of a commercial market [3]. Automatic processing of paper documents is rapidly making importance in India. First research article report on handwritten devanagari characters reported in 1977 but not much research work is reported after that. At present researchers have started working on off-line recognition of handwritten devanagari characters, a number of research reports are available towards devanagari numeral recognition but to the best of our noesis, there are only a few research reports available on off-line handwritten devanagari character recognition after 1977 [3]. To get the idea about advancement and improvement in the OCR's, outcomes of different classifiers and allow for a new benchmark for future research.

In this paper, most desirable feature origin technique HOG, as well as a multiclass model for classification technique used for the classification are mentioned in various segments. This paper is organized as follows. Section 1 reports the introduction part. Section 2 describes the basic properties of handwritten Devanagari characters. Section 3 describes the various literature works for the script. Section 4 proposed method. Section 5describes experimental results and compares the various systems. Conclusion and future scope are provided in section 6. The references include the most relevant papers recently published as well as some older papers, which can give a comprehensive outline of the developments in the field of the research [3].

©IJRASET (UGC Approved Journal): All Rights are Reserved

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887*
*Volume 5 Issue XI November 2017- Available at www.ijraset.com*

## II. PROPERTIES OF DEVANAGARI SCRIPT

Devanagari is one of the most popular script in central or northern India and the most popular Indian language. Hindi is written in devanagari script. Nepali, Sanskrit, and Marathi are also written in devanagari script moreover Hindi is the national language of India. Therefore, the work on Devanagari script is very useful for the country [4].The alphabet of modern Hindi consists of 13 vowels and 34 consonants. The writing style in Devanagari script is from left to right. The concept of upper/lower case is absent in devanagari script. In devanagariscript, a vowel following आand consonant takes a modified shape. The modified shape is placed at the left, right, both or bottom of the consonant. These modified shapes are called modified characters. A consonant or vowel sometimes takes a compound orthographic shape, which we call as compound characters. Compound characters can be combinations of a consonant and a vowel [4]. In that article, we considered 47 basic characters from a self-organized dataset which are shown in figures. he complexity of a handwritten character recognition system increases mostly due to different writing styles [4]. Most of the errors in Hindi OCR's arise due to confusion among the similar shaped characters [11]. In Devanagari, there are many similar shaped characters. Example of some groups of similar shaped characters in fig.1.

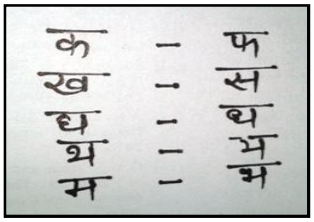



Fig.1Similar shaped characters

devanagari script has 13 vowels which are as shown in fig.2, 34 consonants which are shown in fig.3 [4][5].

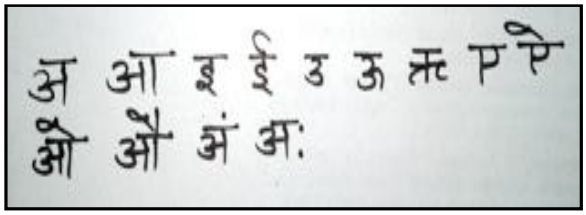



Fig.2 Vowels in Devanagari

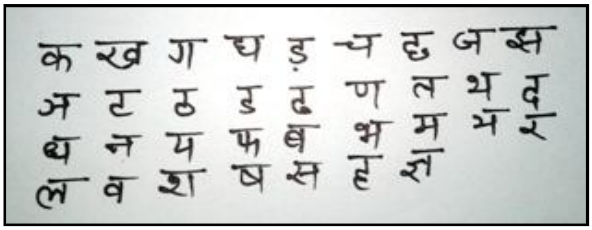



Fig.3 Consonants in Devanagari

A devanagari text line can be partitioned into three zones. The upper zonedenotes the portion above the headline, the middle zone covers the portion between headline and baseline, the lower zoneis the portion below the baseline shown in fig.4.

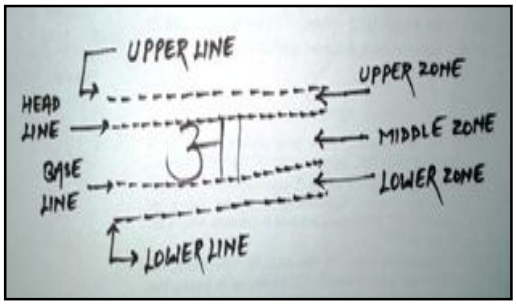



Fig.4 Lines and zones

©IJRASET (UGC Approved Journal): All Rights are Reserved

## III. LITERATURE WORK

In this literature work, we observed that many researchers have done work in the direction to the handwritten Devanagari characters (HDC). The research work on character classification of 9Devanagari script was started in 1970, where Sinha and Maharaja presented a syntactic pattern analysis system for the recognition of devanagari character. The first research on handwritten Devanagari character was published in Seth I. K. and Chattered B. [2] [3]. Researcher started working on the recognition of handwritten Devanagari characters and tried to solve the problem associated with them.

Fuzzy classification based approach for handwritten Devanagari character recognition was proposed by Susana Shelve*etal.*[6]. Recognition system is based on the multi stage classification scheme. The classification stages categorize the characters into smaller groups. The classification is done using two stages, first stage is based on fuzzy inference system and second stage is based on structural parameters. The fuzzy based classification improves the recognition over the crisp classification. It also reduces the burden on the feature extraction and recognition stages to improve the recognition accuracy.

A back propagation neural networks for classification of handwritten devanagari character recognition using wavelet transform is applied to get decomposed images of characters. Statistical parameters are computed over the decomposition to form feature vector, as features for classification proposed by Ad wait Dixit *etal.*[7].

K.V. Kale *et al.*[8], proposed a recognition system for handwritten devanagari Compound Character, based on Legendre moment feature descriptor. Moment function has been successfully applied to many pattern recognition problems. Due to this, they tend to capture global features which make them well suited as feature descriptor. The processed image is normalized to 30X30 pixel size divided into the zone. From this structural as well as statistical feature are extracted from each zone. For classification, they have used Artificial Neural Network.

The Accuracy enhancement of handwritten devanagari character recognition using background elimination and gray level normalization techniques was proposed by Mahesh Jangid*etal.*[9]. The Best choice to extract the features from handwritten Devanagari characters using GLAC (Gradient Local Auto-Correlation) feature extraction technique is used for the experiment. GLAC deals with 2nd order statistics (auto-correlations) which means correlations with neighbor's pixels. Image gradients, in GLAC, are sporadically described in terms of their magnitudes and orientations. All the experiments have done on standard handwritten devanagari characters.

Dinesh V.Rojatkaret al.[10],proposed single hidden layer feed-forward neural network with respect to five-fold cross validation based classification of handwritten devanagari consonant characters. Meticulous experimentation of around seventy-five MLPs showed the overall classification accuracy near to 97% for all classes. This robustness of designed classifier with proposed LRTB features was verified and indicatedthe classifier was robust and perform well for all data partitions.

An algorithm proposed by Mahesh Jaggedet*al.*[11], first to estimate the similar character pairs in devanagari Script and 7 pairs are identified by investigating the confusion matrix. Similar shape characters havea minor difference in shape that's why at the time of recognition (classification) phase, the classifier is being confused with another similar shape characters. This problem can be solved by these timate that minor difference called critical region in the similar shape characters and used the critical region to extract the more features before classification phase. The critical region is estimated by fisher discrimination function. A new kind of masking techniques used to extract the features.

SandyAroraet al.[12],they have proposed a two-stage classification approach for handwritten devanagari characters. The first stage is using structural properties like Shiro Rekha, spine in character and second stage exploits some intersection features of characters which are fed to a feed forward neural network. Simple histogram based method does not work for finding Shiro Rekha, vertical bar (Spine) in handwritten devanagari characters. They designed a differential distance-based technique to find a nearly straight line for Shiro Rekha and spine.

U.Pal *etal.* [13], Represented a Combined use of Support Vector Machines (SVM) and Modified Quadratic Discriminate Function (MQDF) applied for better performance of offline devanagari character recognition with first feature set was computed based on the directional information obtained from the arc tangent of the gradient and curvature-based feature guided by gradient information is computed for the second set of features.

Dayashankar Singh *et al.* [14],they have reported Radial Basis Function (RBF) Neural Network has been implemented on eight directional values of gradient features for handwritten Hindi character recognition. The Radial Basis Function network with one input and one output layer has been used for the training of RBF Network. Experiment has been performed to study the recognition accuracy, training time and classification time of RBF neural network.

©IJRASET (UGC Approved Journal): All Rights are Reserved

## IV. PROPOSED WORK

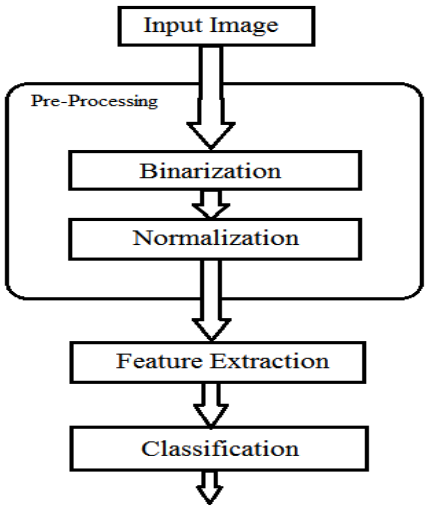



Fig.5 Proposed Method

### *A. Dataset Preparation and Preprocessing*

Datasets is prepared from the handwriting of different persons belonging to the different age groups. In this work we have used basic 47 handwritten Hindi characters as shown in fig.2. The 40 samples of each Hindi character have been written by each person. The main objective behind preprocessing is to remove noise and to make process too simpler. The segmented character is further cropped and then passed to normalization. The cropped characters are of different sizes because the handwritten style of each writer is dissimilar. Normalization is applied on each character to bring all the character in uniform size. These are converted into binary images by choosing such filters. The process of filter and normalization is done by using a preprocessing tool. Then this final processed image passes to the feature extraction process.

### *B. Feature Extraction*

The feature extraction is a crucial part of any classification system. Transforming the input data into the set of features is called as feature extraction. When performing analysis of handwritten Hindi characters data one of the major problem is a number of characters involved. The small set of feature from self-generated dataset that is useful to identifying the pattern in specified classes. Histogram Oriented Gradient (HOG) is used to extract a feature of handwritten Hindi character. Gradient measures the direction and magnitude [19][20] of the maximum variation in intensity in a small zone of each pixel gradients calculated using Sobelkernel. The Sobel kernel used to calculate the horizontal and vertical components of the gradients are shown in fig.6.

| | | |
|---|---|---|
| 1 | 2 | 1 |
| 0 | 0 | 1 |
| -1 | -2 | -1 |

(a)

| | | |
|---|---|---|
| -1 | 0 | 1 |
| -2 | 0 | 2 |
| -1 | 0 | 1 |

(b)

Fig.6. Sober masks for Gradient (a) Horizontal Component (b) Vertical Component.

Scanne dimageis normalized into 20×20 size. The two gradient components at location (i, j) are calculated by the following equations.

Gx = gv (i, j) = f (i−1, j+1) + 2f(i, j+1) + f (i+1, j+1)−f (i−1, j−1) − 2f (i, j−1) – f (i+1, j−1) …… (1)

Gy = gh (i, j) = f (i−1, j−1) + 2f (i−1, j) + f (i−1, j+1)−f (i+1, j−1) − 2f (i+1, j) – f (i+1, j+1) …… (2)

The gradient strength and the direction are calculated as follows:

G (i,j) = $\sqrt{gv\text{^}2(i,j) + gh\text{^}2(i,j)}$…… (3)

©IJRASET (UGC Approved Journal): All Rights are Reserved

$\theta = \arctan(Gy/Gx)$..... (4)

The Horizontal and Vertical templates are used to calculate the gradient components in horizontal and vertical directions, respectively [17][18].

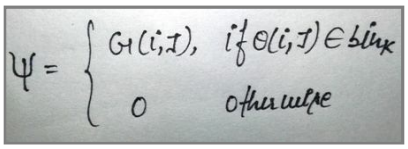


Fig.7 Bins of the Histogram

Bins of the histogram of all blocks, denoted as $bin_k$, can be computed using the block size of the character shown in fig.8. Following equation can be used to compute the feature HOG of each handwritten Hindi character in fig.7.

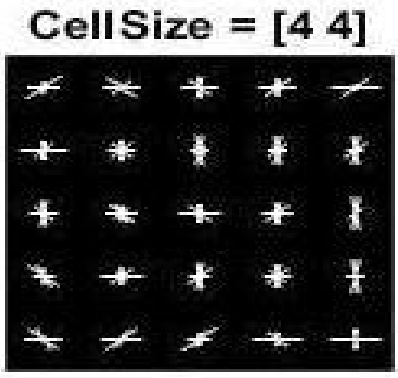



Fig.8. HOG of 4x4 block size of handwritten Hindi Character.

*C. Classification*

Support vector machine for a multiclass classification for higher accuracy, train a binary SVM model or a multiclass error-correcting output codes model (ECOC)[20] containing SVM binary learners to get greater flexibility, use the command-line interface to train a binary SVM model or train a multiclass model composed of binary SVM learners using inbuilt functions. For reduced computation time on high-dimensional data sets that fit in the MATLAB R2016a Workspace, SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class [7]. The best hyperplanefor an SVM means the one with the largest margin between the two classes. Margin refers to the maximal width of the slab parallel to the hyperplane that has no interior data points. The main objective behind this classification system is to achieve good classification accuracy for the identification of handwritten character. The multi-class model for SVM capable of learning to achieve good generalization error-free recognition on these handwritten Hindi character datasets, without any prior knowledge of data, SVM is helpful to achieve robust performance. The concept of an SVM to classify the hyperplane with the maximum margin between the two classes in the feature space, by navigating the input data onto a higher dimensional feature space, which is nonlinearly connected to the input space [9][13]. A maximal margin hyperplane of SVM in feature space is built with kernel function in gene space. Without any computations in the higher dimensional feature space by using kernel functions in the input space, the optimal separating hyperplane can be calculated.

## V. EXPERIMENTAL RESULTS

For implementation Matlab R2016a workspace is used as a tool. The scanned image of handwritten Hindi character is taken as input. Character image set of size 1025 is taken in which625character images are used for training and 400 character images are used for the testing purpose for classification. Firstly, the image needs to be processed so that useful section of the image, on which the classification process will be applied, can be extracted. This is done in preprocessing phase. The scanned image is taken as input and converted into a binary image and further sharpened to remove noise. Then these images are resized into 20x20 dim. HOG is applied on this resized binary image. After applying HOG Feature extraction on the image a vectors generated from an image. For every character feature vector is produced and for classification multiclass model for Support Vector Machine (SVM) is used. When you construct a model for a classification problem you always want to look at the (performance) accuracy of that model as the

©IJRASET (UGC Approved Journal): All Rights are Reserved

number of correct predictions from all predictions made. At the end of the process, a clear way to show the prediction results of a classifier is to use a confusion matrix also called as contingency table. A confusion matrix is a method for summing up the performance of a classification algorithm.

Table-1: Vowels Character Classification

| S.No. | Characters | Results in % |
|---|---|---|
| 1. |  | 96 |
| 2. |  | 100 |
| 3. |  | 98 |
| 4. |  | 94 |
| 5. |  | 98 |
| 6. |  | 100 |

A confusion matrix is formed with the four outcomes produced as result of binary classification these are(TP- correct positive prediction), (FP- incorrect positive prediction), (TN- correct negative prediction), (FN- incorrect negative prediction). This can help in calculating more advanced classification metrics such as precision, recall, specificity and sensitivity of our classifier.

Table-2: Misclassified Characters

| S.No. | Character | Misclassified (%) |
|---|---|---|
| 1. |  | 4 |
| 2. |  | 4 |
| 3. |  | 2 |
| 4. |  | 6 |
| 5. |  | 2 |

In this implementation work, the table shows the confusion matrix in percentage form, the columns of the matrix represent the predicted labels, while the rows represent the known labels. For the given sets in Table-2 following characters are misclassified mostly due to their similar shapes and due to improper character writing styles.

©IJRASET (UGC Approved Journal): All Rights are Reserved

Table-3 Misclassified Characters

| S.No. | Character | Misclassified (%) |
|---|---|---|
| 1. |  | 3 |
| 2. |  | 2 |
| 3. |  | 4 |
| 4. |  | 7 |
| 5. |  | 6 |

Table-4: ClassificationOf Consonants With Different Sample Set

| Sample Set | Accuracy (%) | Specificity (%) |
|---|---|---|
| Set-1 | 99.00 | 99.10 |
| Set-2 | 97.40 | 98.00 |
| Set-3 | 97.60 | 98.10 |
| Set-4 | 99.20 | 99.30 |

Classification accuracy is very important criteriato evaluate the performance of the system. Classification accuracy is the ratio of correct predictions to total predictions made and specificity is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate. The best specificity is 1.0, whereas the worst is 0.0.

Table-5: Comparison Of Classification Accuracy By Other Researchers

| Authors | Data size | Feature Extraction | Classification Method | Results (%) |
|---|---|---|---|---|
| Adwait Dixit et. al. [7] | 2,000 | Wavelet Based Feature | N.N | 70 |
| Sharma et. al. [15] | 11,270 | Chain code | Quadratic | 80.36 |
| Arora et. al. [16] | 1,500 | Combined | MLP | 89.58 |
| DeeptiKhanduja et. al [19] | Not Specified | Statistical features | N.N | 93.04 |
| Pal et. al.[4] | 36,172 | Gradient | MIL | 95.19 |
| Mahesh Jangid et.al. [9] | 36,172 | GLAC | SVM | 95.94 |
| SushamaShelke et.al. [6] | 40,000 | Fuzzy system | FFNN | 96.95 |
| Dinesh V.Rojatkar et. al. [10] | 8,224 | SCG | MLP | 97.00 |
| Proposed Method | 1,025 | HOG | Multiclass model for SVM | 98.32 |

## VI. CONCLUSION & FUTURE SCOPE

We have gone through the procedureof Offline HCR for Hindi Script. In this proposed method a robust feature extraction method has been applied toa self-created dataset. The combination of HOG and multi-class model for SVM classifier provided good accuracy. The overall classification accuracy 98.32% from experimental results we observed that the combination of HOG and themulticlass model for SVM classifier gives better classification accuracy. These methods can be useful for real-time applications like documentation processing, automatic text entry into the computer for desktop publication, processing cheese without human

©IJRASET (UGC Approved Journal): All Rights are Reserved

involvement etc. and theexperimentalresult shows that multiclass model for classifier with global inputs yields good classification accuracy. In India huge volumes of historical documents (handwritten or printed in Devanagarilanguage) remain to be digitized for better access, sharing etc. This will definitely be helpful for other research communities in India in the area of social sciences, economics and linguistics. The digitization of documents and their automatic processing would be easier than keying in the Devanagari text.As we mentioned above we have used robust feature extraction method and multi class model for SVM which gives higher classification accuracy. For further research work, the simplification of HOG using Local binary patterns for gradient and magnitude computation and the complex normalization can be replace by simple linearization, leading to significant saving of area cost without sacrificing too much detection rate and a new SimMSVM algorithm that directly solves a multiclass classification problem. The SimMSVM reduces the size of the dual variables from l*k to l, where l and k are the size of training data and number of classes, respectively. The new SimMSVM approach can greatly speed-up the training process and achieve competitive or better classification accuracies.

## REFERENCES

[1] U. Pal, B.B. Chaudhuri, “Indian Script Character Recognition: A Survey”, by Elsevier Ltd., pattern recognition 37, 1887-1899, 2004.

[2] R. Jayadevan, S.R. Kolhe, P.M. Patil, and U. Pal, “Offline Recognition of Devanagari Script: A Survey”, IEEE Transactions on Systems, Man, and Cybernetics- Part C: Applications and Reviews, vol.41, No. 6, Nov. 2011

[3] Soumen Bag and Gaurav Harit, “A Survey On Optical Character Recognition for Bangla and Devanagari Scripts”, Indian Academy of Sciences, vol. 38, part 1, pp. 133-168, 2013.

[4] U. Pal, T. Wakabayashi, F. Kimura, “Comparative Study of Devanagari Handwritten Character Recognition using Different Feature and Classifiers”, 10th International Conference on Document Analysis and Recognition, 2009.

[5] A.S. Ramteke, M.E. Rane, “A Survey on off-line Recognition of Handwritten Devanagari Script”, International Journal of Scientific & Engineering Research, Volume 3, issue5, May 2012.

[6] S. Shelke, S. Apte, “A Fuzzy Based Classification Scheme for Unconstrained Handwritten Devanagari Character Recognition”, International Conference on Communication, Information and Computing Technology, Jan. 16-17, IEEE, 2015.

[7] A. Dixit, A. Navghane, Y. Dandawate, “Handwritten Devanagari Character Recognition using Wavelet-Based Feature Extraction and Classification Scheme”, Annual IEEE India Conference, IEEE, 2014.

[8] K .V. Kale, S. V. Chavan, M. M. Kari, Y. S. Rode, “Handwritten Devanagari Compound Character Recognition using Legendre Moment an Artificial Neural Network Approach”, International Symposium on Computational and Business Intelligence, IEEE, 2013.

[9] M. Jangid, Dr. S. Srivastava, “Accuracy Enhancement of Devanagari Character Recognition by Grey Level Normalization”, ACM, 2016.

[10] D. V. Rojatkar, K. D. Chinchkhede, G. G. Sarate, “Design and Analysis of LRTB Feature Based Classifier Applied to Handwritten Devanagari Characters: A Neural Network Approach”, IEEE, 2013.

[11] M. Jangid, Dr. S. Srivastava, “Similar Handwritten Devanagari Character Recognition Critical Region Estimation”, International Conference on Advances in Computing Communications and Informatics, Sept. 21-24, IEEE, 2016.

[12] S. Arora, D. B. Charjee, M. Nasipuri, L. Malik, “A Two Stage Classification Approach for Handwritten Devanagari Characters”, International Conference on Computational Intelligence and Multimedia Applications, IEEE, 2007.

[13] U. Pal, S. Chanda, “Accuracy Improvement of Devanagari Character 1Recognition Combining SVM and MQDF”, 2008.

[14] D. Singh, Dr. J. P. Saini, “Hindi Character Recognition using RBF Neural Network and Directional Group Feature Extraction Technique” IEEE, 2015.

[15] N. Sharma, U. Pal, F. Kimura, and S. Pal, “Recognition of offline handwritten Devanagari characters using a quadratic classifier,” in Proc. Indian Conf. Comput. Vis. Graph. Image Process., 2006, pp. 805–816.

[16] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, M. Kundu, and L. Malik, “Study of different features on handwritten Devanagari character,” in Proc. 2nd Emerging Trends Eng. Technol., 2009, pp. 929–933.

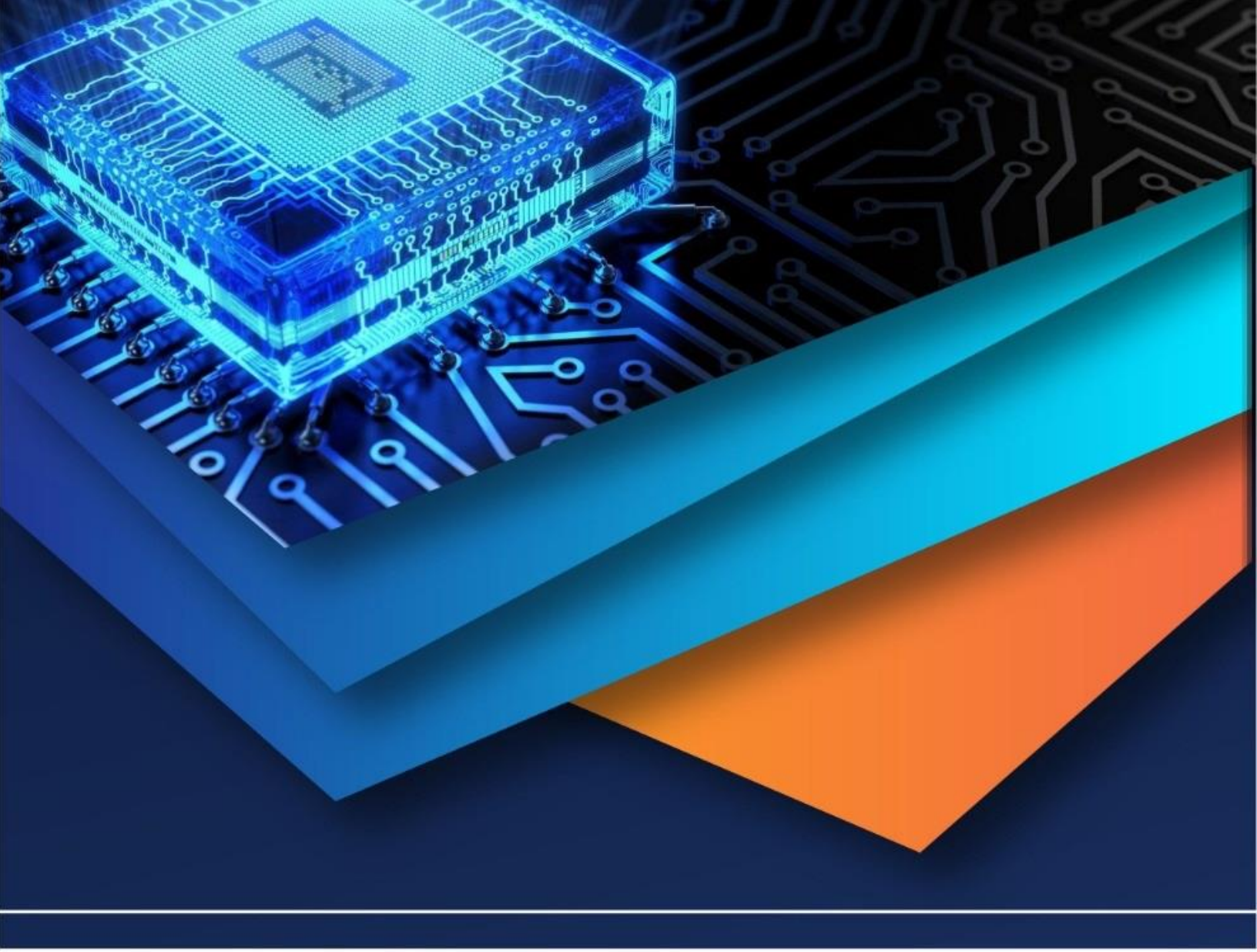
[17] Parshuram M. Kamble, R. S. Hegadi, “Handwritten Marathi Character Recognition using R-HOG Feature”, International Conference on Advanced Computing Technologies and Applications, Procedia Computer Science 45, 266-274, Elsevier, 2015.

[18] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection”, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, Vol. 1, 2005, pp. 886-893.

[19] D. Khanduja, N. Nain, and S. Panwar, “A Hybrid Feature Extraction Algorithm for Devanagari Script”, ACM, 2015.

[20] Mohammad Ali Bagheri, Gholan Ali Montazer, “Error Correcting Output Codes for Multiclass Classification: Application to two Image Vision Problems”, The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, IEEE, 2012.

©IJRASET (UGC Approved Journal): All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)