Combining Multiple Feature Extraction Technique and Classifiers for Increasing Accuracy for Devanagari OCR

Anilkumar N Holambe, Ravindra C Thool

***Abstract*—In this paper we combining statistical, structural, Global transformation and moments features to form hybrid feature vector .We are combining Classifiers for achieving high accuracy for Devanagari Script. To abolish the hitch of misclassification and increase the classifier accuracy, we are combining SVM and KNN together. The dataset used for experiment are created by us.**

***Index Terms*— Devanagari, SVMKNN, Features, Zernike moment.**

I. INTRODUCTION

Pattern recognition is an important area of research. In order to reduce the task of user entering data into a computer system, for this purpose we have to develop optical character recognition (OCR), on the other hand errors are usually found in the recognition process. Therefore, means for finding and correcting such errors in crucial .In order to convert paper world into paper less, many research has been taken place in previous years [1], [2], [3], [4], [5], [6]. The feature extraction is important because we have to find useful information of a character available as an image. On the basis of this information, we have to find the features. This information is quite helpful for taking classification decision and is known as features in pattern recognition terminology. The various feature extraction methods are divided into three categories i.e. statistical, structural, and global transforms [7]. We can combine classifiers in order to get high accuracy in classification of optical character recognition (OCR). The classifiers can be combined using fusion methods using minimum, maximum, average, median, and majority voting is given in [8][9]. Feature extraction and classification are the heart of OCR.

Two specific scenarios are investigated. In the first scenario, the goal is to generate hybrid features. In the second scenario, the goal is to classifier combination approaches for improving Accuracy. Section 2 briefly reviews Devanagari script and Dataset creation, Section 3 Hybrid feature set generation along with Optimal feature set. Section 4 classifier combination Experimental results demonstrating the effectiveness of the new approach are presented in Section 5 and conclusions are offered in Section 6.

II. DEVANAGRI SCRIPT AND DATASET

A. *Review of Devanagri script*

Devanagari is the main script used to write Hindi, Marathi, Nepali and many regional . OCR work on Devanagari script started in 1970s, [10,16].A practice for unnatural hand -printed Devanagari character recognition [11][13][14][15].

B. *Dataset*

We have created different dataset with different ISM software fonts. The font used for this purpose are DVB-DVB-TTBhima,DVB-TTChhaya,DVB-TTDhruv,DVB-TT Dhruv,DVB-TTGanesh,DVB-TTRadhika,DVB-TRaghav,.DVB-TTShridhar,DVB-TTSurekh,DVB-TTYogesh. The font size are 16,18,20,22.Here we have combined numbers, vowels, consonant, it is DS1,second dataset we have created is of Conjunction of Devanagari Consonants (joint characters)with fonts and size this DS2.We have created DS3 from printed document, by segmenting it up to character level.DS4 is created from handwritten character dataset.DS5 of handwritten characters from the segmented documents.DS6 is the data set of words So we six datasets with us for our use. We have arranged the date into the classes. The classes are formed on the bases that each the vowels, consonant, consonant with modifier, number, joint characters are grouped into separate class.

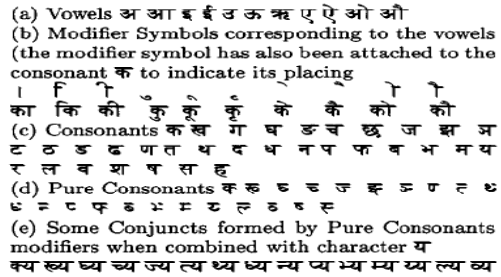


Fig. 1. .Devnagari Vowels, Consonants, Modifier, Conjuncts & Pure Consonants.[17]

Fig. 2. Conjuncts Consonants. [16]

Manuscript received September, 2013.

Anilkumar N Holambe, Shri Guru Gobind Singhji Institute Of Engineering & Technology,Nanded, India.

Dr.Ravindra C Thool. Shri Guru Gobind Singhji Institute Of Engineering & Technology,Nanded, India.

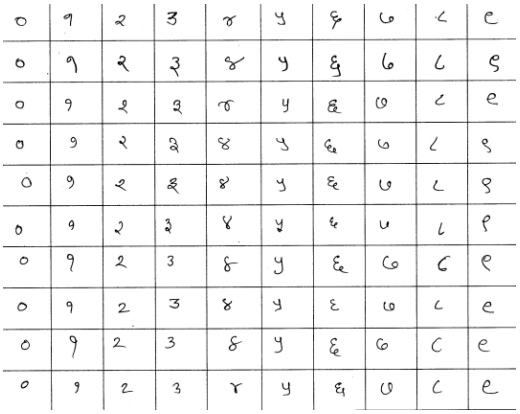


Fig. 3. . Devanagari Numbers

III. FEATURE EXTRACTION AND HYBRID FEATURE FORMATION

Here we are using Holistic approach. Holistic recognition approaches treat words and character as a undividable unit and attempt to recognize words as a whole using their representational features details can be found in [18].In order to reduce error rate we calculate different feature of the same image and combine them so that each and every component of feature should be included. Here we are combining all three types of feature extraction methods that are statistical, structural, Global transformation moments.

A. *Statistical Features*

For statistical features we are developing algorithm base on the method described in [19].

Algorithm

1. Convert the original image matrix into shortest image matrix
2. Divide the image into 9 equal size zones windows i.e. 3x3matrix (features are extracted from each zone rather than on whole image.)
3. Find and collect starting point , intersection point and minor starter location and store in vector.
4. Extract the line segments from the image and store in vector.
5. Forming Direction Vector (Distinguish line as Horizontal, Vertical, Right diagonal, left diagonal line)
6. Find number of any particular line type
7. Find Normalized length of any particular line type
8. Feature Vector formation: The number of horizontal lines, right diagonal lines, Vertical lines. left diagonal lines, intersection point and The total length of horizontal lines, right diagonal lines, Vertical lines. left diagonal lines, intersection point.
9. Store the feature in vector VZN.

B. *Structural Features*

Topological and geometrical properties are used to extract structural features. Now we extract Structural features based on regional properties[24],

a) Euler Number: - It is defined as the difference of number of objects and number of holes in the image.[24]
b) Regional Area: It is defined as ratio of the number of the pixel in the skeleton to the number of pixels in the image.[24]
c) Eccentricity: It is defined as the eccentricity of small ellipse that fills the skeleton of the image.[24]
d) Orientation: The angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second moments as the region.[24]

Store the feature in vector V_{SF}

C. *Global Transformation Moments*

For Global transformation moments features, we use Zernike moments that make the process of recognizing an object scale, translation, and rotation invariant. Here we are using the Zernike moments descried in[20].

"The Zernike moment of order n with repetition m for a continuous image function f(x,y) that vanishes outside the unit circle is

 (1)

For a digital image, the integrals are replaced by summations to get.

 (2)

To compute the Zernike moments of a image, we consider center of image as origin and pixel coordinates are mapped to the range of unit circle."[20]

Therefore $|A_{nm}|$ can be used as a feature. Since $A_{n-m}= A_{nm}$, and therefore $|_{An-m}|=|A_{nm}|$, we will use only $|A_{nm}|$ for features [20]. Store the feature in vector V_{Zr}. We are computing only order 2.So we are getting 4 features.

Finally we will combine all features together in a single vector to form Hybrid feature vector V_{HF}. { $V_{HF}= V_{Zr}+ V_{SF} + V_{ZN}$}

D. *Optimization of hybrid feature vector*

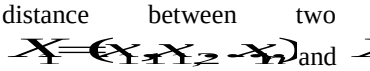
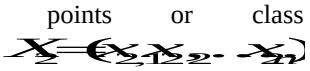
Now we optimize the final vector VHF =k by using following algorithm

1. Select i=1 to N(Class)
2. Train the features with KNN and Classify p random data out of set K. Obtain the efficiency.
3. If the efficiency is low , increase number of features VHF and repeat 2. Else go to step 4
4. Increase the value of p and go to step 3 if p <<K. else go to 5
5. Select the current set of features as the optimum set of features Here we have obtained optimized hybrid features set.

IV. CLASSIFIER COMBINATION

Here we are combining two classifier for improving accuracy. In order to handle large dataset we use SVM Classifier and the rejection and to reduce the error rate we combine KNN with it.

A. *K Nearest Neighbor*

Nearest neighbor classifiers is defined in terms of a distance metric, such as "Euclidean distance"[22]. The Euclidean distance between two points or class  and  obtained from equation 3.

 (3)

The basic steps of the k-NN algorithm are ;

- To compute the distances between the new sample and all previous samples

- To sort the distances in increasing order and select the k samples with the smallest distance values;

B. Support Vector Machine [21]

"The patterns x are embedded, $x_1, x_2, \ldots, x_n \in S$ S can be define as

(4)

The dot product w. x is defined by

$$w \cdot x = \sum_{i=1}^{n} w_i x_i \quad (5)$$

Using the linear classifier defined by the pair (w,b), the class of a pattern x_k is determined with

(6)

Using lagrangian function for optimization and can be formulated into the following expression that represents a linear SVM:

$$w = \sum_{i=1}^{m} \lambda_i y_i x_i \quad (7)$$

After training, we can calculate the class membership for new patterns, different from those used in training. The class of a pattern x_k is determined with

(8)

The classification of new patterns define by the sign of the expression $w \cdot x_k + b$.

(9)

The details of Support Vector Machine and OSH can be found in [21]".

C. Combination of SVM and KNN

Our results show misclassification of class by SVM. Specifically, for class far from the separating hyper plane the SVM classifying algorithm is available, while for samples close to the hyper plane, the KNN Classifying algorithm is suitable. So we are combining both the classifier. Where x_i is support vector. The new classifying algorithm are as follows:

Step 1. If $T_{test} \neq \Phi$ *get* $x \in T_{test}$, If $T_{test} = \Phi$, *Stop;*

Step 2. Calculate $g(x) = sign\left(\sum_{i=1}^{m} \lambda_i y_i x_i \cdot x + b\right)$

Step 3. Compute f(x)=sgn(g(x)). give input to KNN algorithm.

else

Step 4. output =KNN(f(x))

Step 5. T←T-{x} go to step 1.

Here we using RBF kernel .Here we are giving the output of SVM as input to the KNN. The above algorithm is modified from[23].

V. EXPERIMENTS AND DISCUSSION

In our Experiment we used our dataset due to lack of availability of standard database in Devanagari script. We extracted feature set of all the six dataset available with us.V_{HF} is optimized hybrid feature set.

TABLE 1
FEATURE VECTOR SIZE

Datasets	Feature Vector size			
	V_{ZN}	V_{SF}	V_{ZR}	V_{HF}
DS1	81	4	4	89
DS2	81	4	4	89
DS3	81	4	4	89
DS4	81	4	4	89
DS5	81	4	4	89
DS6	81	4	4	89

TABLE 2
KNN CLASSIFIER RECOGNITION RATE IN %

DATASET	FEATURE VECTOR			
	V_{ZN}	V_{SF}	V_{ZR}	V_{HF}
DS1	96.0	93.67	95	97.87
DS2	96.67	94.76	94	97.02
DS3	96.56	94.45	94	97.58
DS4	91.02	90.00	92	92.33
DS5	91.00	89.00	90	90.88
DS6	96.45	94.70	98	97.88

TABLE 3
SVM CLASSIFIER RECOGNITION RATE IN %

DATASET	FEATURE VECTOR			
	V_{ZN}	V_{SF}	V_{ZR}	V_{HF}
DS1	96.56	91.00	94.00	94.20
DS2	96.00	94.76	96.00	94.75
DS3	94.87	94.07	95.88	96.23
DS4	84.78	83.01	84.00	84.77
DS5	83.00	81.00	83.00	84.00
DS6	96.34	95.55	97.00	97.88

TABLE 4
SVM KNN CLASSIFIER RECOGNITION RATE IN %

DATASET	FEATURE VECTOR			
	V_{ZN}	V_{SF}	V_{ZR}	V_{HF}
DS1	98.78	97.02	97.89	98.34
DS2	98.67	96.60	96.78	98.23
DS3	98.56	97.01	97.00	98.78
DS4	92.08	91.00	92.23	92.75
DS5	92.55	91.06	92.80	92.88
DS6	98.78	97.86	98.45	98.88

From above table of result we observe that SVMKNN has high accuracy. We found that KNN has high complexity and classification time than SVM. But SVM can classify large data but Classification error is there around the hyper plane vector.SVMKNN algorithm work on large set of data here SVM efficiently uses the KNN algorithm features for High accuracy.

VI. CONCLUSION

A KKN and SVM individual give better result on numbers ,single characters means characters without modifiers, and with joint characters and character with modifiers the error rate is more and time for recognition is high. But SVMKNN algorithm gives better result on all type of Devanagari data sets.

REFERENCES

[1] C.Y. Suen, M. Berthod, and S. Mori, "Automatic Recognition of Hand printed Characters – The State of The Art," Proc. IEEE, vol.68, no. 4, pp. 469-487, Apr. 1980.

[2] J. Mantas, "An Overview of Character Recognition Methodologies," Pattern Recognition, vol. 19, no. 6, pp. 425-430, 1986.

[3] V.K. Govindan and A.P. Shivaprasad, "Character Recognition – AReview," Pattern Recognition, vol. 23, no. 7, pp. 671-683, 1990.

[4] S. Mori, C.Y. Suen, and K. Yamamoto,"Historical Review of OCRResearch and Development," Proc. IEEE, vol. 80, no. 7, pp. 1029-1058, Jul. 1992.

[5] H. Bunke and P.S.P. Wang (eds.), Handbook of Character Recognition and Document Image Analysis, World Scientific Publishing, Singapore,1997

[6] N. Nagy, "Twenty Years of Document Image Analysis in PAMI,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no.1, pp. 38-62, Jan. 2000.

[7] D. Trier, A. K. Jain, T. Taxt, "Feature Extraction Method for Character Recognition - A Survey", Pattern recognition, vol.29, no.4, pp.641-662, 1996

[8] Ludmila IK, "A Theoretical Study on Six Classifier Fusion Strategies," IEEE Trans. On Pattern Analysis and Machine Intelligence, v. 24, No. 2, pp. 281-286, Feb. 2002.

[9] J Kittler, M Hatef, RPW Duin, and J Matas,"On Combining Classifiers," IEEE Trans. On Pat. Analysis and Machine Intel., Vol. 20, No. 3, Mar. 1998. Mar. 1998.

[10] R.M.K. Sinha, H. Mahabala,, "Machine recognition of Devanagri script", IEEE Trans. System, Man Cybern. 9(1979) 435-441.

[11] Plamondon, R. Srihari, S.N. ,Ecole Polytech.,Montreal, Que. , " Online and Offline HandwritingRecognition : A comprehensive Survey,1EEE Transactions On Pattern Analysis And Machine Intelligence. VOL. 22, NO. 1. JANUARY 2000

[12] U. Pal , B.B. Chaudhuri , "Printed Devanagri script OCR system", Vivek 10 (1997) 12-24

[13] S. Palit, B.B. Chaudhuri, "A feature-based scheme for the machine recognition of printed Devanagri script", P.P. Das, B.N. Chatterjee (Eda.) Pattern Recognition, Image Processing and Computer Vision, Narosa Publishing House: New Delhi, India 1995, pp. 163-168

[14] I.K. Sethi, B. Chatterjee, "Machine recognition of constrained hand-printed Devanagri numerals", J. Inst.Electron. Telecom. Eng. 22 (1976) 532-535.

[15] I.K. Sethi, B. Chatterjee,"Machine recognition of constrained hand-printed Devanagri characters", Pattern Recognition 9 (1977) 69-76

[16] R.M..K. Sinha, "A syntactic pattern analysis system and its application to Devanagri script recognition", Ph.D. Thesis , Electrical Engineering Department, Indian Institute of Technology, India, 1973.

[17] V. Bansal and R. M. K. Sinha, "Integrating Knowledge Sources in Devnagri Text Recognition," IEEE Transactions on Systems, Man, & Cybernetics Part A: Systems . the International Society for Optical Engineering. v 3964 2000. p305-312

[18] S. Madhvanath and V. Govindaraju. , " The role of holistic paradigms in handwritten word recognition. ,"IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 23(2):149-164, 2001.

[19] M. Blumenstein, B. K. Verma and H. Basli, , "A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters,"7th International Conference on Document Analysis and Recognition (ICDAR '03) Eddinburgh, Scotland: pp.137-141, 2003.

[20] Alireza khotanzad and Yaw hua hong, "Invariant Image Recognition by Zernike Moments,"IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 12. No. 5. Pp489-497 may 1990

[21] Ovidiu Ivanciuc, "Applications of Support Vector Machines in Chemistry", In: Reviews in Computational Chemistry, Volume 23, 2007, pp. 291–400.

[22] Tao Hong, Stephen W. Lam, Jonathan J. Hull and Sargur N. Srihari, "The Design of a Nearest-Neighbor Classifier and Its Use for Japanese Character Recognition", in the Proc.of the thied Int. Conf.on Document Analysis and Recognition,Aug.1995,270-273.

[23] Rong Li, Hua-Ning Wang, Han He, Yan-Mei Cui and Zhan-Le Du, " Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting," Chin. J. Astron. Astrophys. Vol. 7, No. 3, 441–447-2007

[24] http://www.mathworks.com/help/images/ref/regionprops.html