THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Improving Similar Looking Characters Recognition In Gurmukhi Script Using Spectral Overlap and Compensation Over Projection Histogram

Tarun Jangwal[1],

[1]Assistant Professor, School of Computer Science and Engineering, Lovely Professional University, India. Tarun.24044@lpu.co.in

Abstract

Achievement rates of the Optical Character Recognition (OCR) algorithms of published Gurmukhi Script documents are fairly remarkable with state-of-art, levels of accuracy ranging from 85 to 95 percent for specific. Nonetheless, additional enhancement of this level of accuracy is required to support actual applications. One of the bottle necks is known as similar looking characters in further improving the accuracy. In this paper, we outline the similar looking characters in Gurmukhi Script and creation of a specialized classifier for these closely matching characters. State-of-art OCR's output is used and the similar looking character set groups are further provided to an advanced classifier to improve precision. Support vector machine (SVM) algorithm is used to make this classifier and it uses feature vectors taken from near matching characters' spectral coefficients of projection histogram signals.

Keywords *- OCR, Gurmukhi Script, Pattern Classifications, Feature Extraction, Similar looking Character*

I. INTRODUCTION

In this era of computing internationalization, all the languages of the world are converted into its virtual script. The success depends on the language and user base, of course. Nonetheless, the initiatives are limited to languages that have a critical mass of user base because of the minimalist of the market context. Punjabi is the Punjab state's official language is an instance of such a scenario. Punjabi (what is spoken) is a language. Gurmukhi is a script (the way Punjabi is written). Punjabi is an Indian language, a part of the same Indo-European language group, belonging to the outer circle of Indo-Aryan languages and distantly related English.

Gurmukhi script language has been well incorporated with digital technology at a very basic level. For desktop publishing applications, unicode fonts for Gurmukhi Script are eligible. All non-Unicode fonts and Unicode fonts are commonly obtainable, although there is a range of keyboard support which needs to be streamlined for use in a practical sense. To varying degrees of success, further technologization such as speech recognition, simulation, and identification of character is ongoing. At the Punjabi University, research and development efforts are being made. Through this paper, we propose some reasonable improvements throughout established OCR systems by mounting our standardized pattern classifier on similar looking characters in Gurmukhi Script, thereby providing a way for Gurmukhi Script to enhance the precision character recognition of OCR systems.

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

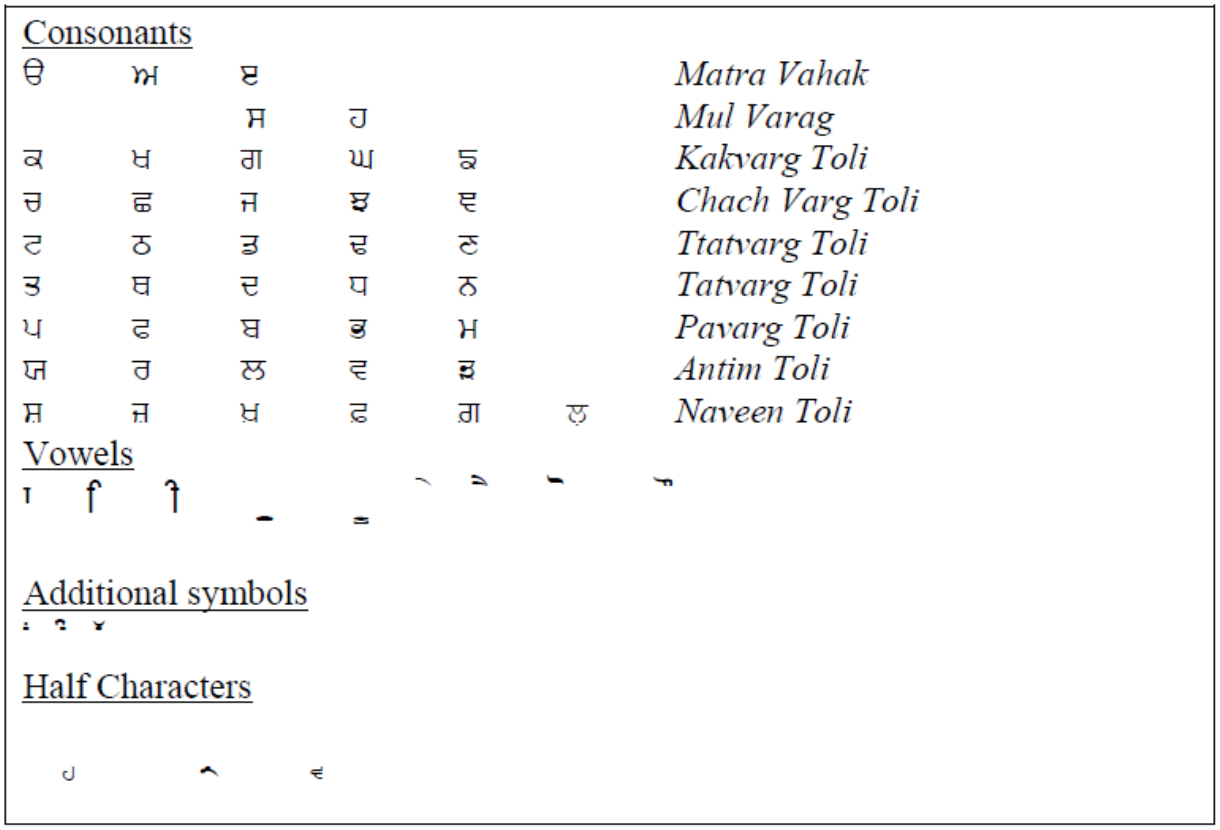


Fig 1 : Gurmukhi Character Set

Close-matching characters							
ਚ	ਰ	ਜ	ਜ਼	ਸ	ਸ਼	ਪ	ਮ
ਧ	ਯ	ੳ	ੳ	ਥ	ਬ	ਗ	ਗ਼
ਭ	ਭ	ਅ	ਘ	ਫ	ਫ਼	ਖ	ਖ਼

Figure 2 some similar looking characters in Gurmukhi Script

Gurmukhi script alphabet consists of 41 consonants, 12 vowels and 3 characters in the form of half characters are present in the feet of characters as shown in Figure 1. Writing style is from left to right. The concept of upper/lower-case characters is absent in Gurmukhi. Figure 2 shows some of the similar looking characters in the Gurmukhi Script character set that we considered in this study. For the characters to be integrated into the standard OCR system a specialized classifier system is proposed as given in Figure 3.

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

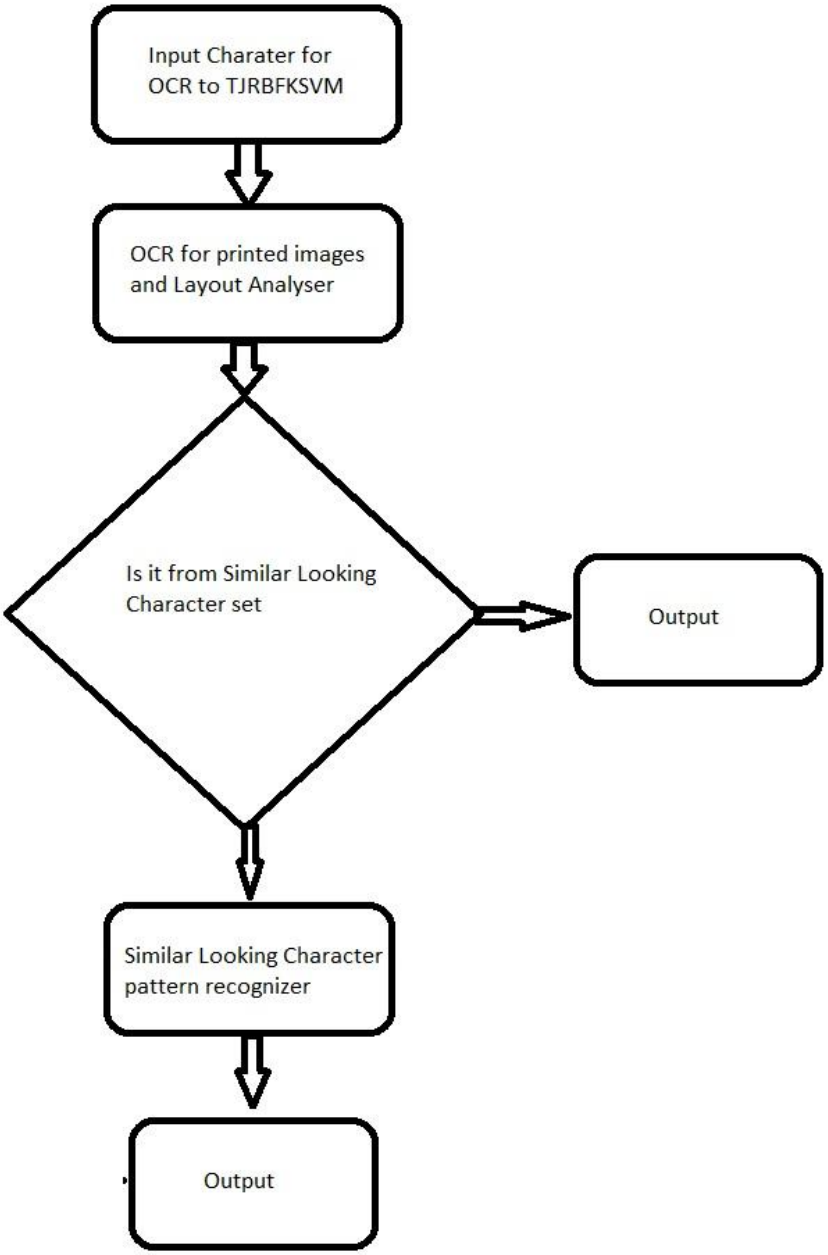


Figure 3 The proposed Similar looking specialized classifier System

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

II. Gurmukhi Script OCR Systems

We concisely define Gurmukhi Script's state-of-art OCR program that is identical to other languages as well. This is divided into three phases: preprocessing, identification and post-processing. First the text file is scanned then after scanning, a bit mapped object will be generated in the preprocessing portion. Then black and white binary images are generated from gray-level images, and then generic noise reduction algorithm is implemented, then if required, in the image blur corrections are also done. Segmentation of the site and description of the section are also done in order to infer that character segmentation is there[3]. This section of the preprocessing cycle is considered an evaluation of the format. When object segmentation is completed, the pattern recognition process is effective for characters. The character code is defined for every input character, in this character recognition process features are derived from a character learning package that uses a professional method. The features are extracted and then fed into this classifier for each new incoming character that generates a draft message. The recognizer's heart is the combination of realistic elements reflecting the nature of the character's individuality. It is very important to fine tune pattern recognition systems, with help vector machines or related soft computing approaches. To increase the level of accuracy even after the classifier's acceptance some post validation can be performed in different of ways. To identify some ordinary unmanageable character blend that may come in predictor, the character errors can be approved by making use of the language's linguistic rule. This is a stretched out method to improve the entire system's reliability. Nonetheless, this does not guarantee that it will work in every situation as there can be mistakes in the identification of characters that may not break any language law. The instance in English might be "for" and "if."

State-of-the-art Gurmukhi Script OCR solutions use smart character recognition or study of topological properties. We focus the appreciation on character-specific features. To order to identify individual characters, basic characteristics used are such as points, vertical lines and closed loop etc. To derive these attributes from segmented characters, the character recognition element is loaded into a binary tree classifier[5]. The character is identified after passing through the classifier's dissimilar stages and the corresponding character code is assigned.

The basic character classification uses the stroke-based and topological features [6]. Our factors in choosing the functionality include character design, function quality, font and width freedom, as well as the rate of removal of the app. They divided their characteristics into three groups. Form 1 designs contains of horizontal bottom lines, left bar, vertical right bar, vertical mid bar, vertical left line, and right row. Type 2 characteristics consist of aspect ratio, loops, loop height, loop size, loop location, vertical cutting points and horizontal cutting points.

Initial classification of characters use the stroke based and topological features [6]. Size independence, speed of feature extraction, font, character structure, and reliability of the feature, these are considered in selecting the features. There is three type of feature grouping. First group features horizontal bottom bars, horizontal bar, right part bar, vertical mid bar, vertical left bar, vertical right bar and left part bar. Second group has aspect ratio, loops, loop height, loop size, loop location, vertical cutting points and horizontal cutting points. Third group features of high importance they are features based on numbers, number end points and position of these end point. First and second group and are thought to be independent of style and font size, hence defining most important characteristics.

The binary tree classifier used to equate received data using variety of features in hierarchical classifier. By determining the categorizing based on separability of classes the selection of features is done. Using decision tree method computation period is reduced. However, the system's accuracy depends on design of the decision tree. This is restrictive factor for maximum accuracy achieved through this paper and

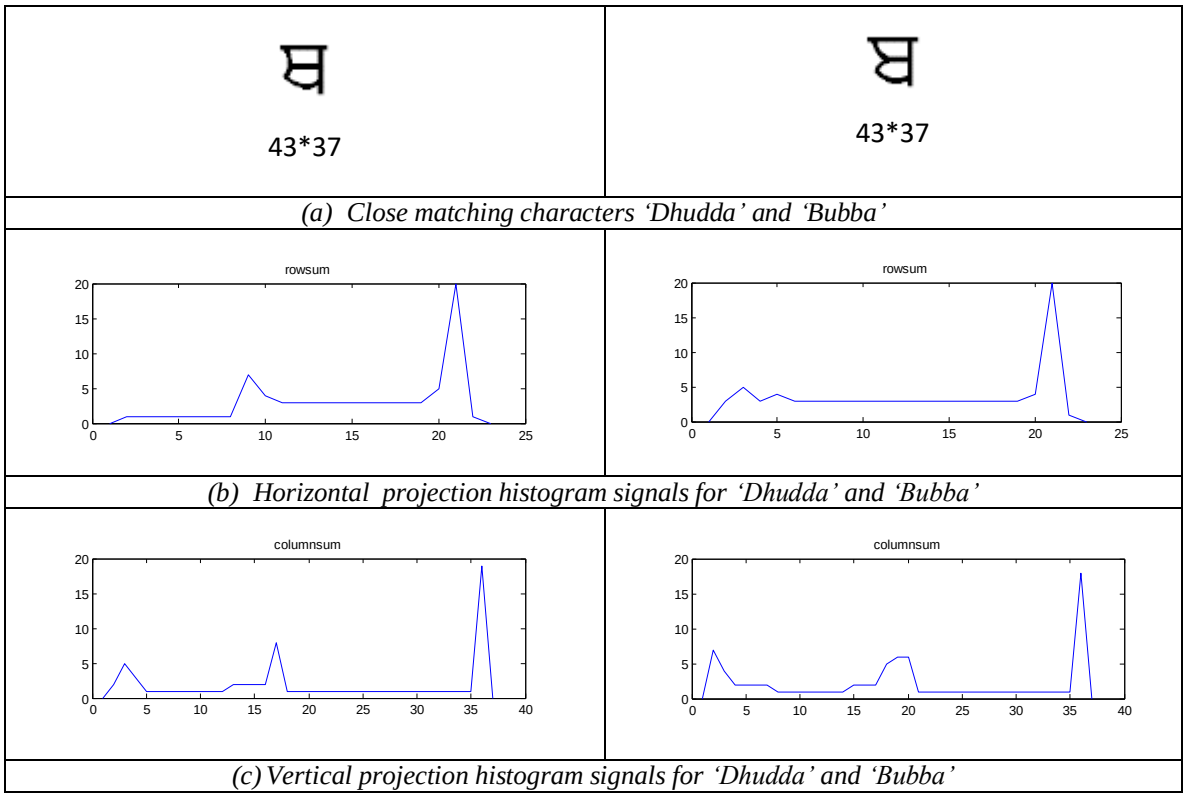
Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

therefore research described in our paper we discarded with a binary tree classifier and instead used dynamic computational method.

III. FEATURE VECTOR FOR SIMILAR LOOKING CHARACTERS

In Gurmukhi Script, we find the near matching character collection (Figure 2) and define a classifier dependent on the exclusive extraction of features. Extraction of features is a crucial stage in the development of some identification of patterns. The objects that cannot be taken directly to the classifier as input data are to be recognised character images that are often gigantic. It is therefore important to select a limited set of powerful features that will loyally reflect the individuality of the item to be classified. Goal is high rate of recognition for which we have to derive some feature. Different handwriting or fonts may result in a high degree of variation, It also needs to be taken into consideration when choosing the. Characteristically, extraction of features is based on three forms of features global transformations, structural and statistical. Statistical feature is the portrayal through statistical distribution of points of a character image and takes care to some degree of design variation[7]. Zoning, projections & profiles and crossings & distances are the key statistical features used for character portrayal. Structural features also enhanced acceptance of irregularities and differences in form. Such style of depiction can also teach some information about the object's structure or provides understanding about what kind of component are used in creating a object. Structural features are based on geometric and topological character attributes, such as cross points, aspect ratio, loops, branch points, strokes and directions two-point inflection, bottom or top horizontal curves, etc. Moments and Global Transformations render identification phase as rotation invariant and scale translation. From moment coefficients original objects may be wholly restored.



(a) Close matching characters 'Dhudda' and 'Bubba'

(b) Horizontal projection histogram signals for 'Dhudda' and 'Bubba'

(c) Vertical projection histogram signals for 'Dhudda' and 'Bubba'

Figure 3. character's extracted Projection Histogram Signal– for example of *'Dhudda' and 'Bubba'*. Although the characters are Similar-looking, they have very exclusive signals.

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Nevertheless, in the case of near matched characters, the use of the above features has not been shown to be adequately effective as demonstrated by the recorded accuracy levels. The designers have been pointing out that the confusion created by similar looking characters in Gurmukhi Script is one of the triggers of dive in accuracy rates. The remark is has inspired us for the study and analysis.

By selection of each character's features and creating a decision tree. We carefully considered the character images which ignore character type or character content and tried to extract unique feature. A powerful and unique trait vector is derived by use projection histogram. A projection histogram is to be attained by measuring a character image's pixel distribution horizontally and vertically. Figure 3, displays two near corresponding characters' projection histogram signals of Gurmukhi Script.

A projection histogram's basic conversion is to translate the initial one-dimensional signals from 2-D signals. Such signal are deformation and noise autonomous, yet they rely on the rotation[6]. Discrete Fourier Transforms are used in digital signal processing devices [5], one dimensional signal can also be interpreted quickly. Fourier analysis the recognized statistical approach through which a large number of constituent sinusoidal signals can be evaluated as a quantity of any dynamic signal[2]. Applying Discrete
A collection of n-point information can be translated to the coefficient vector's n-point DFT. It is well established that it is possible to remove the large coefficients in these types of vector and reconstruct the signal having superior resemblance. Its certainly our endeavour to be within reach of extraction of a feature vector. Vector T(n) is used to represent N-point projection histogram signal. Using the following equation, the spectral content is determined:

$$T(k) = \sum_{z=0}^{Z-1} T(n)\, e^{-j2\pi kz/Z}, \text{ where } k = 0, 1, 2, .., Z-1$$

By dumping upper coefficients z-point vector T(k) may be decreased to any size of choice. Its admissible because in determining character visual identification, high frequency variations in representations of characters are not required. Let Tv and Th be image's vertical and horizontal histogram projection signal, then Vector J a feature can be derived as:

$$J = [f_{h1}, f_{h2}, ..., +f_{hM};\ f_{v1}, f_{v2},,.... f_{vM}]$$

Where f_h r is is the horizontal projection histogram image's r-th Fourier coefficient and fvr is the vertical projection histogram image's r-th Fourier coefficient The image dimensions are considered as ZxZ, where $M \leq Z$.

IV. RESULT, DISCUSSION AND CONCLUSION

We used the training data-set for similar-looking character recognizer. We then split the similarly corresponding character training data-set of the whole data-set and then divided it to two equivalent test data-sets and training data-sets along projected histograms derived in 2d in the X/Y direction. Coefficients of the Discrete Fourier Transform (DFT) are determined and other half is discarded and generated a feature vectors for every similar character set. To develop the classifier tools we used a Support Vector Machine (SVM)[9].

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Traditional artificial neural networks that are available are used in SVMs n-dimensional plane separation enhancement. Maximum band of separation among pattern cluster is obtained as an alternative to the simple linear separation and therefore SVMs are considered better than other current soft computing approaches. To build our framework, we used the TJRBFKSVM software[8]. We used the RBF Kernel because of its ability and flexibility for handling nonlinear class-separation boundary. Once we tested our trained SVM we received performance rates in the same range as of the system. It means similar-looking character recogniser's performance is better than the original system. Switching to a subjective feature extraction from an objective signal processing based feature extraction may likely be applied to full data-set of characters. If more precision is required, we believe image can be converted into a better illustration that exibits the ability to have highly influential objective feature extraction. A nominee we are currently investigating for such transformations is the RQA (Recurrence Quantification Analysis). We also feel that the scope of the OCR in Gurmukhi Script needs to be reopened in order for moving from text documents to character identification from captured images and images from video has not yet been addressed.

.

REFERENCES

[1] G.S. Lehal, Nivedan Bhatt, "A Recognition System for Devnagari and English Handwritten Numerals", *Proc. ICMI, Springer*, pp. 442-449, 2000.

[2] Reena Bajaj, Lipika Dey, Shantanu Chaudhuri, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers", *Sadhana* Vol. 27, Part 1, pp. 59-72, February 2002.

[3] U. Bhattacharya, B. B. Chaudhuri, "A Majority Voting Scheme for Multiresolution Recognition of Handprinted Numerals," *Seventh International Conference on Document Analysis and Recognition (ICDAR)* - Volume 1, pp. 16, 2003.

[4] U. Bhattacharya, S.K. Parui, B. Shaw, K. Bhattacharya,"Neural Combination of ANN and HMM for Handwritten Devanagari Numeral Recognition", *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[5] G. S. Lehal and Chandan Singh, "A Gurmukhi script recognition system", *Pattern Recognition, Proceedings, 15th International Conference*, Barcelona, Spain, 3-7 Sep. 2000, vol. 2, pp. 557-560.

[6] Manish Kumar, RK Sharma, and G. S. Gehal, *Degraded Text Recognition of Gurmukhi Script*, [Online]. Available: http://hdl.handle.net/10266/465, Accessed: 10 Jan. 2012.

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

[8] R.J. Ramteke, S.C. Mehrotra, "Recognition of Handwritten Devanagari Numerals", *International Journal of Computer Processing of Oriental Languages, Chinese Language Computer Society & World Scientific Publishing Company*, 2008.

[9] G.G. Rajput, S.M. Mali, "Fourier Descriptor Based Isolated Marathi Handwritten Numeral Recognition", *International Journal of Computer Applications(IJCA)*,Vol. 3, No. 4, June 2010.

[10] Apurva A. Desai, "Gujarati Handwritten Numeral Optical Character Reorganization through Neural Network", *Pattern Recognition*, Vol. 43, Issue 7, pp. 2582-89, July2010

Copyright © 2019Authors