INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

POSSIBLE USE OF OCR FOR RECOGNITION OF KORKU LANGUAGE TEXT

ARVIND ARJUNRAO TAYADE, ANAND MUKUNDRAO KHANDARE

Assistant professor in Computer Science, G.S. Science, Arts and Commerce College, Khamgaon.

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract*:* In this paper we have describe the possible use of OCR to create the technique for korku language text recognition. Several methodologies used for creating techniques which are belonging to OCR family. Some feature extraction techniques are used in this mechanism. Pattern recognition techniques are also implemented for recognition of korku language text.

Keywords: OCR, Segmentation Recognition

Corresponding Author: MR. ARVIND ARJUNRAO TAYADE

Access Online On:

www.ijpret.com

How to Cite This Article:

Arvind Arjunrao Tayade, IJPRET, 2016; Volume 4 (9): 888-892

PAPER-QR CODE

INTRODUCTION

Under the Devanagari script, the huge of languages and semi languages are offended. In some cases the pronunciation is different i.e. the writing in same script but in case of reading there are two words having same meaning. For example sene in korku and ja in Marathi that means go in english. In the case of writing phase of korku people, write with both languages in same script then challenging task is how to recognize that language and translated in the original or major language i.e. Marathi language which is understandable to major majority people. To create this type of mechanism firstly we need to analyze the document find each text and matching with database and find out its corresponding text.

The methods popularly used in the early stage of OCR (optical character recognition) research and development are template matching and structural analysis [1]. Optical Character Recognition (OCR) is one of the earliest applications of pattern recognition. In OCR, the characters or symbols to be recognized can be have machine printed characters or handwritten characters. Machine recognition of handwritten characters continues to be a topic of interest among many researchers, primarily due to the potential applications in various fields such as – document recognition, cheque processing, form processing, postal address recognition, etc.

The communities in India are linguistically diverse. They speak not only the scheduled but also non-scheduled languages and dialects. The people of India linguistically are divided into different language families, such as Indo-Aryan, Dravidian, Austro-Asiatic, Tibeto-Berman & And amanese. The present study about the Korku language belongs to Austro-Asiatic language family, mainly spoken by the Korku tribes of Maharashtra and Madhya Pradesh. The word Korku means simply ‘men’ or ‘tribesmen’, koru being their term for a man and ku a plural termination [2].

In this paper, we have work on the possible use of OCR in the korku language text recognition and how to translate in the Marathi for the use of efficient manipulating the document having korku and Marathi.

KORKU LANGUAGE DICTIONARY

To solve the problem of understanding, after matching the text there will be need of dictionary. This dictionary contains the meaning of Korku language text in Marathi script. When we arrange the dictionary then maintain some parameter regarding to this dictionary. Numbering must be in proper order due to accuracy of that system. In this paper we have work only the recognition domain but in overall system depend only on dictionary.

II. METHODOLOGY

OCR:

There are various recognizer are available in market like MICR , Magnetic stripe, Bar code, Optical Mark Reading etc. This data storage technique only machine dependent i.e. the format of that can't be readable to human being but in case of Optical Character Recognizer technology, the format of documentation is readable to human being and also machine. This technology is unique, it does not require control of process that produce the information [3].

When we go through OCR technology it will consist several stages. This is the easiest way to interact the machine with human languages following figure no 1 shows the various stages of OCR system.

A. Optical scanner

This is a sensing process in this process, document of recognizable will scan through optical scanner which is in gray level\scale image. After sensing image applies the thresholding process on that image. In this process, image gets its threshold i.e. maximum and minimum value. Simply we can convert image in binary form [3]. That process help to next step segmentation of input image.

B. Location Segmentation

Segmentation is mechanism for separating the each character from the text. Simply this is a decomposition tool of image. Segmentation is a process that determines the constituents of an image [3]. Those will identity the regions of document where is data printed. From the previous process, we take the binary image for segmentation. Challenging task for segmentation is identifying shirorekha from the devnagari script and some attach word. To resolve this problem, we are work with row-wise and column-wise addition of binary digit. In case of horizontal segmentation, highest addition means shirorekha. In same manner horizontal segmentation low addition means separation of characters. Following Figure-2 shows the actual mechanism of segmentation i.e identifying addition and then segment the each character from word.

C. Preprocessing:

At the time of image sensing there are various noises are entered in the image due to its resolution. Some noises are implemented through different intensity or illumination. When the image contains noise, recognition is complicated, due to artifacts. If noises on character then that can't treated as its original meaning, because of number of characters having minor changes then identification processes is so critical. To solve those problems, preprocessing is most important task in this process. Various filtering techniques are implemented on the image

for smoothing process. In other point of view the size of character, positioning and other parameter are measure in preprocessing process. Following figure-3 shows the smoothing of image.

D. Feature Extraction:

This is the heart of OCR technology. Without extraction of feature, we can't be match with other sample. In feature extraction terminology there are various ways for extracting feature from preprocessed image. There are mainly three categories for extracting the feature [3].

Structural Analysis.

Transformations and Series Expansions.

The Distribution of Points.

We are go through transformation and series expansions feature extracting technique. But our main vision to identify the text of korku language so we can extract whole word from document and match with our database. If it is matched then that korku language text gets translated in Marathi language script.

E. Recognition and Post processing:

To recognize the extracted features with standard trained database, numbers of techniques are available. In which classification method, structural method are available and some statistical method also use for reorganization. We are go through structural method because recognition of whole text is depending on its structural shape. Output of the whole process is in individual character format. To identifying the errors, false acceptance rate from the output images, post processing is a needed task.

After going through this methodology, we have get the korku language text. Assurance of trusted image then next methodology implemented on that image and translated in to conventional Marathi. This is readable to all kind of person who knows Marathi language.

Following figure-4 shows the overall terminology regarding to technique.

CONCLUSION:

In this paper, we are only focus on the use of OCR for recognition of korku language text. By using this technique there will be a effective use of OCR to information society and also some backward people who doesn't know Marathi or other mostly used language. The problem of communication with small community and its documentation will efficiently identified and understandable to citizens who does not know korku but know Marathi. This is also useful for interact their ideas with its own language. For future work, this technique identifying handwritten document of Korku language text. In This paper we have work only on primary level of research work. So some lacunas can be presented in this research work.

REFERENCES:

1. S. Mori, C.Y. Suen, K.Yamamoto, Historical review of OCR research and development, Proc. IEEE, 80(7): 1029-1058, 1992.
2. Russell, R.V. & Hiralal. 1916. The Tribes and Castes of the Central Provinces of India (London: Mac Milan and company, rpt. 1975, Delhi: Cosmo Publications), vol. III, pp. 550-70.
3. Line Eikvil Optical Character Recognition December 1993.
4. Veena Bansal and R. M. K. Sinha Integrating Knowledge Sources in Devanagari Text Recognition System, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 30, NO. 4, JULY 2000.
5. Swapnil Khedekar Vemulapati Ramanaprasad Srirangaraj Setlur Venugopal Govindaraju Text - Image Separation in Devanagari Documents, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03) IEEE 0-7695-1960-1/03 2003.
6. Srirangaraj Setlur Suryaprakash Kompalli Vemulapati Ramanaprasad Creation of data resources and design of an evaluation test bed for Devanagari script recognition IEEE 0-7803-7868-7/03 2003.
7. U. Bhattacharya, S. K. Parui and S. Mondal Devanagari and Bangla Text Extraction from Natural Scene Images ,10th International Conference on Document Analysis and Recognition 2009.