Text Document Analysis and Recognition

[1.]Chandralekha Padvekar , [2.]Pranoti Shukla, [3.]Neha Shah, [4.]Ketkee Sonawane
[1,2,3,4.]BE-Computer Engg
Smt.KashibaiNavale College of Engg,Smt.kashibaiNavale College of Engg.
Pune, IndiaPune,India

***Abstract*-Our purpose is to create an application that allows user to jump to the correct section of an audio book by taking a picture of the page they are reading in a paper book. We used image processing algorithms such as the Hough transform to rotate pictures of the page, Otsu's method to threshold and binarize the image and brick wall coding (BWC) to detect features on the page. We obtained 40% accuracy over a dataset of 20 pages. In future work we could improve our accuracy by using angles between features which are more invariant to the pictures the user takes.**

Keywords:ImageProcessing,OpticalCharacterRecognition,Pattern Recognition,HandwritingRecognition, Audio Conversion.

I. INTRODUCTION

The Here in our application we are going to focus on the act to make our application portable to be used by the people who are in a hurry or cannot read. Old age persons cannot read a lot of the times so they can use our application to convert the text into audio and listen to it whenever they want. The user does not need to have a technical knowledge to use this application. To achieve this we are going to make available different languages in which the user wants to listen and the text will be provided by our application if he wants to send it to another person.

This is an offline (does not require internet connection to work) Optical Text Recognition application which can be used to convert text (English) on a paper to editable digital text on your device. Extracted/converted text can be listen in audio format.

In terms of image processing, our algorithm must extract the specially defined descriptor of taken image (word lengths or locations, neighbors or possible geometric features could be used). The same feature extraction algorithm should be applied to the all images in the database. Then it must map and evaluate these features via pairwise matching. Timestamp of the matched page must be accurately computed so that the correct timestamp is selected with an acceptable error probability.

II. LITERATURE SURVEY

A. RELATED WORK DONE

Existing system {CHARACTER RECOGNITION IN NATURAL IMAGES}, tackled the problem of recognizing characters in images of natural scenes. In particular, it focused on recognizing characters in situations that would traditionally not be handled well by OCR techniques.It presented an annotated database of images containing English and Kannada characters. The database comprises of images of street scenes taken in Bangalore, India using a standard camera. The problem was addressed in an object categorization framework based on a bag-of-visual-words representation. It assesses the performance of various features based on nearest neighbour and SVM classification. It is demonstrated that the performance of the proposed method, using as few as 15 training images, can be far superior to that of commercial OCR systems. Furthermore, the method can benefit from synthetically generated training data obviating the need for expensive data collection and annotation.

In more recent work in the same area, Lopresti and Zhou evaluated the performance of several classical and enhanced IR models using simulated OCR data. To enhance traditional IR models to deal with the imperfect data, they used approximate string matching and fuzzy logic. In general, they were able to show that the new methods are more robust to noisy data than the original methods, suggesting that simple enhancements can be used to improve performance. Ohta et al, described a system for full text search in which they augment three probabilistic text retrieval methods with knowledge about expected OCR errors. The approach used confusion information for specific characters, along with bi-gram probabilities of character occurrences to create multiple possible search terms for each initial search term. After performing the search with each new term, the validity of returned documents is based on the confusion and bi-gram occurrence probabilities. The results claim increases from 2-3% in recall with decreases or 4-5% in precision. Fujisawa and Marukawa used a similar approach in which they use confusion statistics to generate an enhanced finite state machine for query terms in Japanese text. An alternative approach to attempting to modify the query to deal with poor quality is to modify the matching algorithm, as described by Takasu in. To obtain speed, the approach uses a two-stage algorithm where the first stage uses a fast string matching algorithm to generate match candidates, and the second stage uses a more appropriate similarity distance measure, such as the Levenshtein distance. As with this approach shows slight improvements in recall, with slight decreases in precision. Many documents which are created electronically have both structured and unstructured components. For example, electronic mail (Email) delineates the sender, receiver, date and subject, in addition to the message. It is useful to distinguish between document indexes which rely on objective, structured identifiers, such as authors' names, titles and publishers, and non-objective identifiers which are extracted directly from the text content. If the document analysis front end provides objective identifiers, standard database operations can be used to query document databases. In the absence of objective identifiers, methods for characterizing the full text content of the converted documents must be developed. The latter is a more challenging problem, however, and where most research is being carried out.

Research in IR has led to the use of a wide variety of techniques for automatic indexing. Historically, experts have been called upon to manually provide a concise index of content descriptors for each document. More recently, techniques such as inverted indexes, term weighting and its variations, and the extraction of relationships between terms to preserve content, have evolved to provide advanced automatic in-dexing. For retrieval, vector space, probabilistic and boolean retrieval models provide a foundation for document similarity. The basic idea behind most approaches to text indexing and retrieval is to provide the ability to characterize the text corpus in a meaningful way, to allow users to provide a query as a set of terms, and to provide mechanisms to retrieve, in ranked order, the most relevant documents for that query.1 One common way of characterizing a document's content is to consider the full text, filter out common \stop" words which have a negligible effect on the content, and then represent the document by a term vector consisting of the frequencies of meaningful terms. Furthermore, can be stripped (stemming), low-frequency words replaced with thesaurus equivalences, and high- frequency words replaced with phrases, in an attempt to reduce inappropriate variability in the text.

III. System features

A.Specification

It may happen that user wishes to read many pages in one go, this feature is provided by the systemAlso, one of the significant feature of the system is that it can be operated in offline mode .Thus,the user need not have the internet connection in order to use the system once the application is downloaded.

This systemfocuses on extracting features from images of physical book's pages and convert them into audio.

B.Functional Specification

System allows user to take snap shot using camera. It doesn't ask user for any type of authentication. user can take more than one snap shot of physical book's pages or any handwritten document. User should take the snap shots of multiple pages in sequence in order to read them in continuous manner. System provides multiple language support for audio. System supports only English language data to get extracted from image. System also supports Email/SMS facility for extracted text.

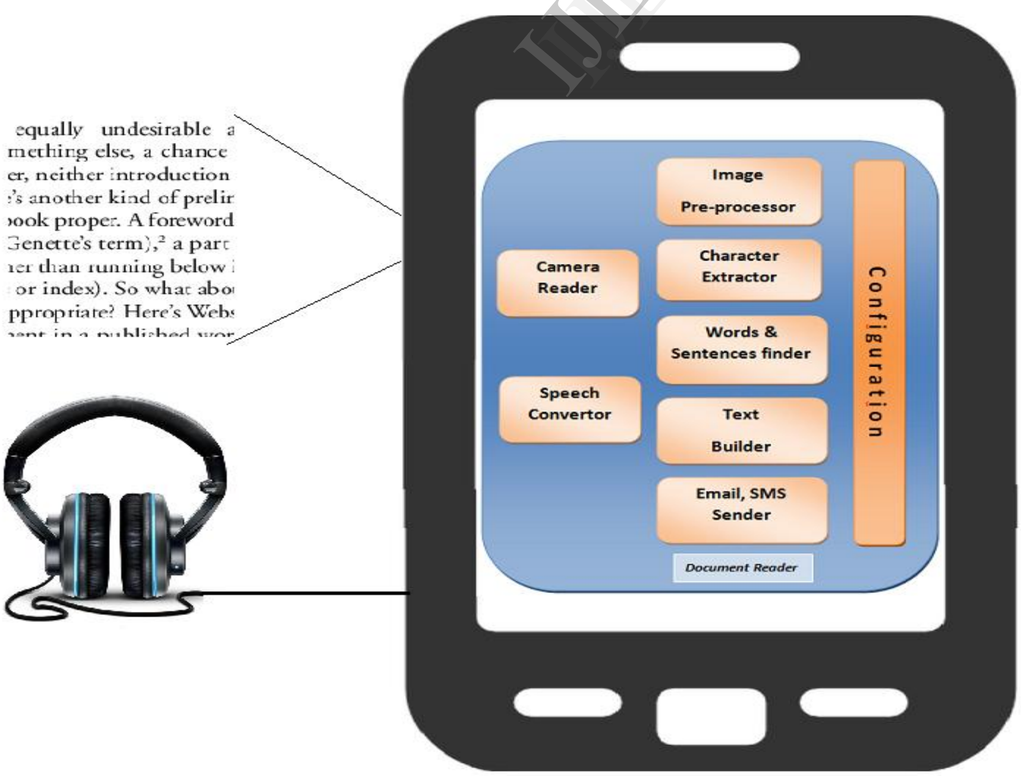

Fig 1: System Architecture

IV. ARCHITECTURE

A. System Architectural design

In the above architectural diagram fig. 1, system consists of 5 modules which are mainly designed to handle image to speech conversion. Those 5 main modules are:

a. Image Pre-Processor

Image pre-processor carries out image binarization, in which it the given image into gray scale.it then extracts the relevant part of the binary image taking into consideration the average letter width height estimation.

b. Character Extraction

From the gray scale image, the character extractor extracts the characters from the text taking into consideration the default white space between each character or word.

c. Word and Sentence Finder

A set of training images of alphanumeric characters is given as input to the system which maps the extracted character from the gray scale image to the matched character in training images and forms respective words.

d. Text Builder

Incase ofawordmismatch thesystem triestofindthenearestword existinginthedictionarywiththe helpofWordnet[3][4].For example,ifaword‘*Amorphous*’isminedas ‘*Amorfous*’which doesn’texistsinthedictionary thenthesystempredictsand returnsthebestmatch(es)as‘*Amorous*’and‘*Amorphous*’ by carefully assessing thesubstrings.Sinceitispossibletohave ProperNounsandwordswhichdonotresideinthe dictionary, themismatchwordsarenotauto-corrected andincaseofone ormorepredictions theuserisaskedtopickthebestone manually. Ifthesystemfailstoproducethebestmatchesor doesn’tproduceamatchatallthenithighlightsthosewordsfor supplementarycontemplationbytheuser.

e. Speech Conversion

The sentences which are build are converted to audio format using google API. And this audio format file can be further transferred via email or sms.

V. PROJECT FLOW

The overall flow of the system goes throughthese stages as shown in fig 2.

.

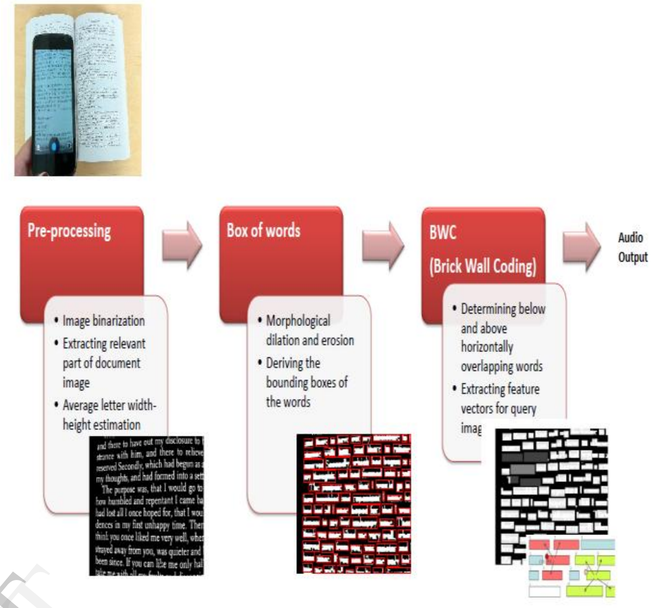
Pre-processing
Image binarization
Extracting relevant part of document image
Average letter width-height estimation
Box of words
Morphological dilation and erosion
Deriving the bounding boxes of the words
BWC
(Brick Wall Coding)
Determining below and above horizontally overlapping words
Extracting feature vectors for query imag
Audio Output

Fig. 2:Project Flow diagram

VI. CONCLUSION

The system presents a novel approach towards converting printed and handwritten text to audible media. This application enables users to get relevant information from the books and other documents without going through the tedious work of reading them. As the demand for systems which can recognize omnifont is increasing, this system will prove to be an efficient tool for recognizing unconstrained print.

In the future, system can be improved with the advancements in the field of Image Processing and Speech Synthesis. In this world where technology evolves everyday and decreasing computational restrictions open up, there seems no end to the development of new methodologies for character recognition.

ACKNOWLEDGMENT

This research was supported by SMT. KASHIBAI NAVALE COLLEGE OF ENGINEERING, SINHGAD TECHNICAL EDUCATION SOCIETY, and PUNE University the authors also wish to express their gratitude towards Computer Department of SKNCOE for their continuous support on this research.

REFERENCES

[1] AnjumAli,MahmoodAhmad, NasirRafiq,JavedAkber,Usman Ahmad andAhahwarAkmal,“*LanguageIndependentOptical Character RecognitionforHandWrittenText*,”inMultitopicConference,2004. Proceedings ofINMIC 2004. 8th International, 2004, pp. 79-84.

[2] RamanathanR.,PonmathavanS.,ValliappanN., ThaneshwaranL.,Nair A.S.,SomanK.P.,"*Optical CharacterRecognitionforEnglishandTamil UsingSupportVectorMachines*,"inAdvancesinComputing,Control,&Telecommunication Technologies, 2009. ACT'09.International Conference,2009, pp. 610– 612.

[3]GeorgeA.Miller(1995),“*WordNet:ALexicalDatabasefor English*,” in Communicationsof theACMVol.38,No.11:3941.

[4] Christiane Fellbaum(1998, ed.), “*WordNet: An ElectronicLexical Database*,”Cambridge,MA:MITPress.

[5] Kazutaka Takeda, Koichi Kise, and Masakazu Iwamura. , “*In Document Analysis and Recognition* (*ICDAR*),” 2011 International Conference on, pages 1054_1058, 2011.