International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 11 | Nov -2016 www.irjet.net p-ISSN: 2395-0072

BAYESIAN CLASSIFICATION FOR MULTI-ORIENTED AUDIO TEXT RECOGNITION SYSTEM

[*1]Mrs. Vennila S, [*2] Ms. Mahalakshmi J.

[*1]Assisant Professor, Department of Computer Science Auxilium College (Autonomous), Vellore, TamilNadu, India

[*2] PG Student, Department of Computer Science Auxilium College (Autonomous), Vellore, TamilNadu, India

---

Abstract - ***A text to Audio converter convert's normal language text into Audio. Text to Audio converter is useful in different applications. Customer support dialog systems Interactive voice response (IVR) systems etc and are also useful in an applied research. This application is more helpful in banking, toys and many other applications like checking marks, railways, aid to the physically challenged persons, language education and fundamental and applied research etc. But text to Audio conversion is not that much easy for machine as it is for human. Basic steps that machine has to follow for text to Audio analysis are database creation, character recognition and text to Audio conversion. This paper surveys methods related to character recognition as well as approaches used for text to Audio conversion for machine we are going to develop an on-line Audio-to-text engine. However, the transfer of Audio into written language in real time requires special techniques as it must be very fast and almost 100% correct to be understandable. The objective of this review paper is to recapitulate and match up to different Audio recognition systems as well as approaches for the Audio to text conversion and identify research topics and applications which are at the forefront of this exciting and challenging field.***

Key Words: *Interactive voice response (IVR), Text to Audio, Audio-to-text engine.*

I. INTRODUCTION

A Text-To-Audio (TTA) synthesizer is a computer-based system that should be able to read any text aloud, whether it was directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. Let us try to be clear. There is a fundamental difference between the system we are about to discuss here and any other talking machine (as a cassette-player for example) in the sense that we are interested in the automatic production of new sentences. This definition still needs some refinements. Systems that simply concatenate isolated words or parts of sentences, denoted as Voice Response Systems, are only applicable when a limited vocabulary is required (typically a few one hundreds of words), and when the sentences to be pronounced respect a very restricted structure, as is the case for the announcement of arrivals in train stations for instance. In the context of TTS synthesis, it is impossible (and luckily useless) to record and store all the words of the language. It is thus more suitable to define Text-To-Speech as the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter.

Automated text classification, also called categorization of texts, has a history, which dates back to the beginning of the 1960s. But the incredible increase in available online documents in the last two decades, due to the expanding internet, has intensified and renewed the interest in automated document classification and data mining. In the beginning text classification focused on heuristic methods, i.e. solving the task by applying a set of rules based on expert knowledge. This approach proved to be highly inefficient, so nowadays the focus has turned to fully automatic learning and clustering methods.

The task of text classification consists in assigning a document to one or more categories, based on the semantic content of the document. Document (or text) classification runs in two modes:

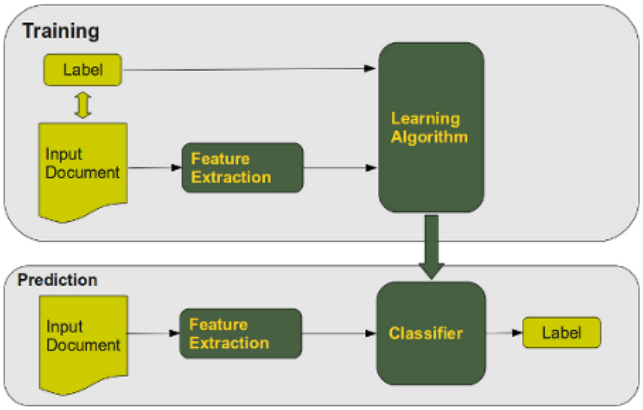

Fig 1.1: The training phase and the prediction (or classification) phase.

1.1 Optical Character Recognition

Optical character recognition usually abbreviated OCR, is the mechanical or electronic translation of images of and written, type written or printed text. optical character recognition belongs to the family of techniques performing automatic identification. For character recognition authors used different methods that are stated below.

Characters in a text are of different shapes and structures. Text extraction may employ binarization or directly process the original images it consist a survey of existing techniques for page layout an analysis. Mathematical morphology is a topological and geometrical based approach for image analysis. It provides powerful tools for extracting geometrical structures and representing shapes in many applications. Morphological feature extraction techniques have been efficiently applied to character recognition and document analysis, especially if dedicated hardware is used.

They proposed an algorithm for text extraction based on morphological operations.

OCR which is the acronym for Optical Character Recognition. This technology allows a Machine to automatically recognize a character through an optical mechanism.OCR is the process of translating scanned images of typewritten text into machine editable information. If we read a page in language other than our own, we may recognize the various characters, but be unable to recognize words. However, on the same page; we are usually able to interpret numerical statements-the symbol for numbers is universally used.

II. METHODOLOGY

Text to speech system has two parts namely natural language processing and speech synthesis (digital signal processing).

2.1 Natural Language Processing (NLP)

NLP produces phonetic transcription together with prosodic feature of the input text. In this TTS system, NLPcomprises of three main components such as text analysis, phonetic conversion and prosodic phrasing.

2.1.1Text Analysis

In this TTS system, the input sentence is segmented into token. After tokenization, each word is determined as part of speech (POS) tagging. Part-of-speech is a process assigning correct POS tag to each word in a sentence from a given set of tags. Bigram Model is used for POS tagger. This method is to perform POS Tagging to determine the most likely tag for a word, given the previous and next tags [3]. This can be calculated by using equation (1).For Bigrams, the probability of a sequence is just the product of conditional probabilities of its Bigrams. So if $t1$, $t2 \ldots tn$are tag sequence and $w1$, $w2$,.....wn are corresponding word sequence.

$$P(ti \mid wi) = P(wi \mid ti) \cdot P(ti \mid ti+1)$$

Where ti denotes the tag sequence and wi denotes the word sequences. P $(wi \mid ti)$ is the probability of current word given current tag. Here, P $(ti \mid ti+1)$ is the probability of a current tag given the previous tag. This provides the transition between the tags and helps capture the context of the sentence.

2.1.2 Phonetic Conversion

In this system, Dictionary based approach is used for phonetic transcription of input word. So, any type of input text that does not include in the dictionary cannot run.

2.1.3 Prosodic Phrasing

Prosodic Phrasing is to assign the phrase of the input text .In this part, chink n'chunk prosodic phrasing, is used. In this model, word classes are identified into chink and chunk group .Then, input words are compared with chink or chunk group. Prosodic phrase break is automatically set when a word belonging the chunks groups. This method basically corresponds to function and content word classed, with some minor modification.

2.2 Speech synthesis

The speech synthesis is to produce speech as natural and intelligible sound .There is many methods in speech synthesis. Among these, concatenative speech synthesis is natural in comparison with other methods. In this TTS system, sub-types of concatenative synthesis such as unit selection speech synthesis, phoneme based speech synthesis and domain specific synthesis are applied.

2.2.1 Unit Selection Speech synthesis

This algorithm selects an optimum set of acoustic units from the speech database to match with the given phoneme stream and target prosody. A selection mechanism using two cost functions – target cost and concatenation (join) cost is applied to find the best sequence of units [4] .The target cost function typically consists of several subcomponents of phonological features such as identity of its context, positional features and numerical features such as phrasing. The target cost, $Ct\ (ti \mid ui)$ can be computed by the following equation (2)

$$Ct\ (ti \mid ui) = wj\ q\ tj{=}1\ Cjt\ (tt \mid ui)$$

Where p represents the number of the target cost Components, $wj\ t$ and is a feature weight of the j-th component. The concatenation or joint cost function accounts for the acoustic matching between pairs of consecutive candidate units and it can be calculated by using equation (3)

$$Cc\ (ui{-}1 \mid ui) = wj\ q\ cj{=}1\ Cjc\ (ui{-}1 \mid ui)$$

Where q represents the number of the concatenation cost components The unit selection module is to find the speech unit sequence which is described in equation (4)

$$C1^{n} = \min C(t1^{n}, u1^{n})$$

The selection of the optimal speech unit sequence incorporates a Viterbi search.

2.2.2 Phoneme based speech synthesis

Phonemes are the small pieces of speech unit. English language has about 44 phonemes of which 22 sounds are vowels and 22 sounds are consonants [5]. Phoneme based speech synthesis is the concatenation of phonetic units to form word. . Using phonemes as the synthesis unit requires a small storage, but it causes little discontinuity between adjacent units.

2.2.3 Domain-specific Synthesis

Domain-specific synthesis concatenates pre-recorded words and phrases to create complete utterances. The technology is very simple to implement .The level of naturalness of these systems can be very high because the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings.

III PROBLEM DESCRIPTION

3.1. Segmentation

Segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters which are recognized individually. Usually this segmentation is performed by isolating each connected component that is each connected black area. This technique is easy to implement, but problems occurs if characters touch or if characters are fragmented and consist of several parts. The main problems in segmentation may be divided into four groups:

- Extraction of touching and fragmented characters.
- Distinguishing noise from text.
- Mistaking text for graphics or geometry

3.2. Pre-processing

Pre-processing techniques are binary document containing text and/or graphics. In character recognition systems usually applications utilize grey or binary images since processing that is computationally high and tedious task. Such type of images may also contain non-uniform background and making it difficult to extract the document text from the image without performing some kind of pre-processing, therefore, the desired result from pre-processing is a binary comprising text only. Thus, to achieve this, several steps are needed, such as image enhancement techniques to remove noise or improve contrast in the text, and Thresholding to remove scenes or noise, third, page segmentation to isolate graphics from text, fourth, character segmentation to separate characters from each other and, finally, morphological processing to enhance the characters in cases where Thresholding and/or other pre-processing techniques eroded parts of the characters or added pixels to them.

3.3. Feature Extraction

It is the most important module implemented in OCR system. Feature extraction is a technique to extract the main features of a character to recognize it properly with maximum accuracy. It includes steps like cropping and scans lines. Feature extraction process is followed by Recognition process. It Compares character with the available templates.

3.4. Recognition

Recognition process Compare character with the available templates matching techniques is different from the others in that no features are actually extracted. Instead the matrix containing of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype is computed, and the class of the prototype giving the best matches assigned to the pattern.

3.5. Text to Audio conversion

Using Speech Application Program Interface (SAPI) the given text is converted into audio. The Speech Application Programming Interface is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within Windows applications. To date, a number of versions of the API have been released, which have shipped either as part of a Speech SDK, or as part of the Windows OS itself.

IV ALGORITHM IMPLEMENTATION

SpyNB learns user behavior models from preferences extracted from click through data. Assuming that users only click on documents that are of interest to them, SpyNB treats the clicked documents as positive samples, and predict reliable negative documents from the unlabeled (i.e. un clicked) documents. To do the prediction, the “spy” technique incorporates a novel voting procedure into Naive Bayes classifier to predict a negative set of documents from the unlabeled document set. The details of the SpyNB method can be found in. Let P be the positive set, U the unlabeled set and PN the predicted negative set (PN ⊂ U) obtained from the SpyNB method. SpyNB assumes that the user would always prefer the positive set over the predicted negative set.

Abstractly, the probability model for a classifier is a conditional model

$$p(C|F_1,\ldots,F_n)$$

Over a dependent class variable C with a small number of outcomes or *classes*, conditional on several feature variables F_1 through F_n. The problem is that if the

number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable.

Using Bayes' theorem, this can be written

$$p(C|F_1,\dots,F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$$

In plain English, using Bayesian Probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features F_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$p(C, F_1, \dots, F_n)$$

Which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned} p(C, F_1, \dots, F_n) &= p(C)\ p(F_1, \dots, F_n|C) \\ &= p(C)\ p(F_1|C)\ p(F_2, \dots, F_n|C, F_1) \\ &= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ p(F_3, \dots, F_n|C, F_1, F_2) \\ &= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ \dots p(F_n|C, F_1, F_2, F_3, \dots, F_{n-1}) \end{aligned}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$, given the category C. This means that

$$\begin{aligned} p(F_i|C, F_j) &= p(F_i|C), \\ p(F_i|C, F_j, F_k) &= p(F_i|C), \\ p(F_i|C, F_j, F_k, F_l) &= p(F_i|C), \end{aligned}$$

and so on, for $i \neq j, k, l$. Thus, the joint model can be expressed as

$$\begin{aligned} p(C|F_1,\dots,F_n) &\propto p(C, F_1, \dots, F_n) \\ &\propto p(C)\ p(F_1|C)\ p(F_2|C)\ p(F_3|C)\ \cdots \\ &\propto p(C) \prod_{i=1}^{n} p(F_i|C)\,. \end{aligned}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is:

$$p(C|F_1,\dots,F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

where the evidence $Z = p(F_1,\dots,F_n)$ is a scaling factor dependent only on $F_1,\dots,F_n$, that is, a constant if the values of the feature variables are known.

4.1 OCR Algorithm

Input:

Complex network, whose state x evolves according to

$dx/dt = F(x)$

Initial state x_0

Desired target state, x*

Constraints on allowed perturbations:

$G(x_0, x_0') < 0$

$h(x_0, x_0') = 0$

Output:

x0', state in the targets basiyan attraction for x0-x0' an eligible perturbations

SPEECH SYNTHESIS

A Text-To-Speech (TTS) synthesizer is a computer based system that should be able to read any text aloud, when it is directly introduced in the computer by an operator. It is more suitable to define Text-To- Speech or speech synthesis as an automatic production of speech, by 'grapheme to phoneme' transcription.

A grapheme is the smallest distinguishing unit in a written language. It does not carry meaning by itself. Graphemes include alphabetic letters, numerical digits, punctuation marks, and the individual symbols of any of the world's writing systems. A phoneme is "the smallest segmental unit of sound employed to form meaningful utterances" The first task faced by any TTS system is the conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme to phoneme conversion. This consist different methods followed by different author that are summarized as follows:

A. Chauhan,Vineet Chauhan,Surendra P. Singh, Ajay K. Tomar, Himanshu Chauhan uses

The basic types of synthesis system the following are:

- Formant
- Concatenated
- Prerecorded

A. Concatenative Synthesis:

Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech.

V SYSTEM DESIGN

5.1 System Structure

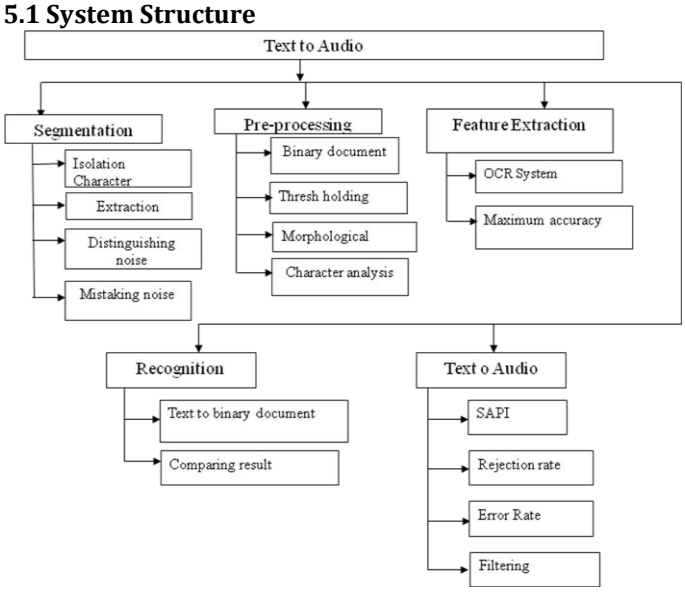

Fig5.1 System Structure

5.2 System Architecture

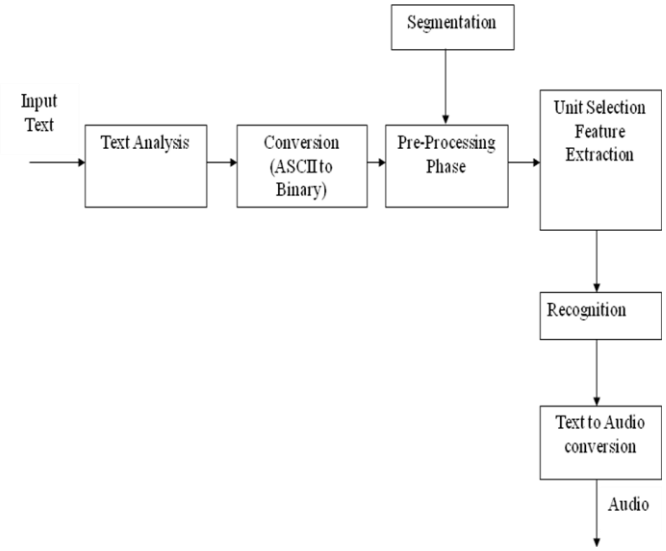

Fig 5.2 System Architecture

5.3 System Processing Diagram

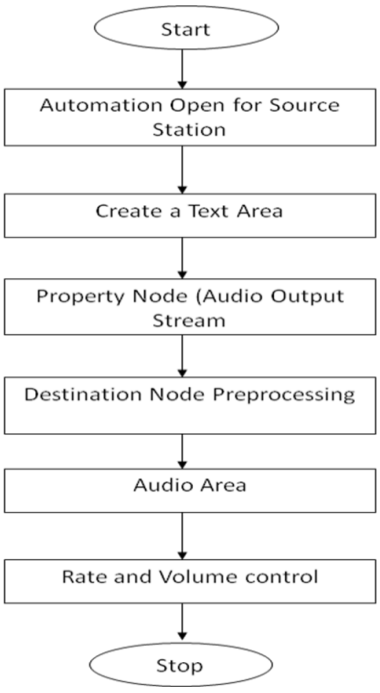

Fig 5.3 System Processing Diagram

VI. EVALUATION RESULT:

Software implementation is based on Dot Net programming language. In domain specific synthesis, the input number (one or more digits) can be pronounced speech easily and quickly. The output speech is natural and intelligible like human speech. But the domain specific synthesis is not general purpose. For the combination of words; it must be preprogrammed to synthesize speech.

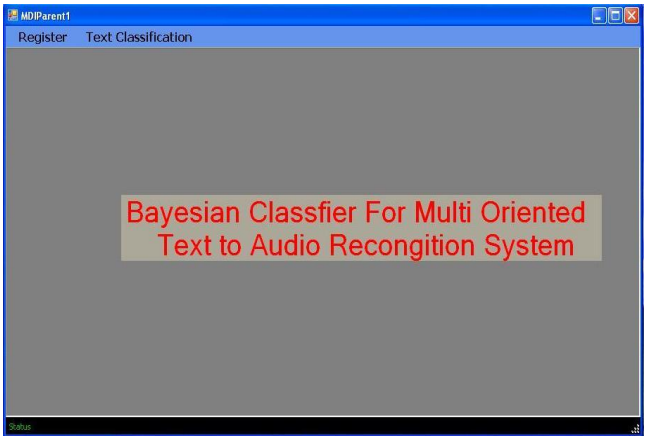

Fig: 6.1 Home Page

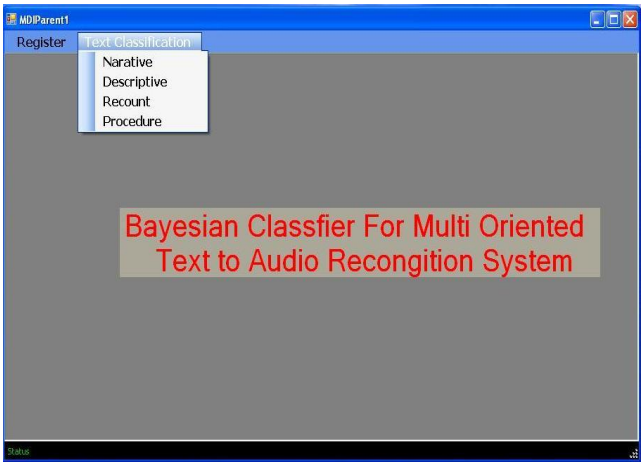

Fig: 6.2 Text Classification

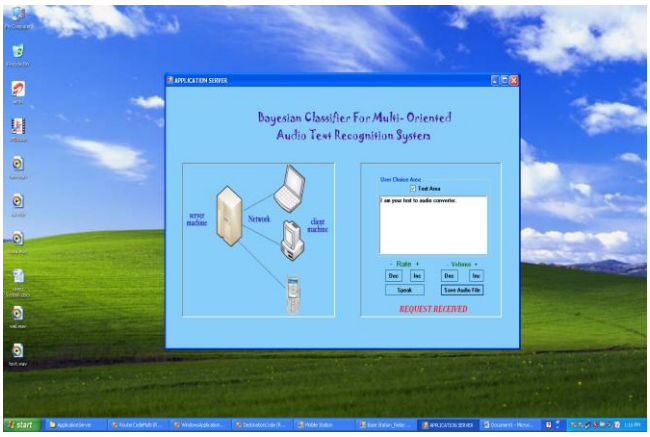

Fig: 6.3 Application Serve to Text area

Fig: 6.4 Mobile Station

Fig: 6.5 Convert To Audio

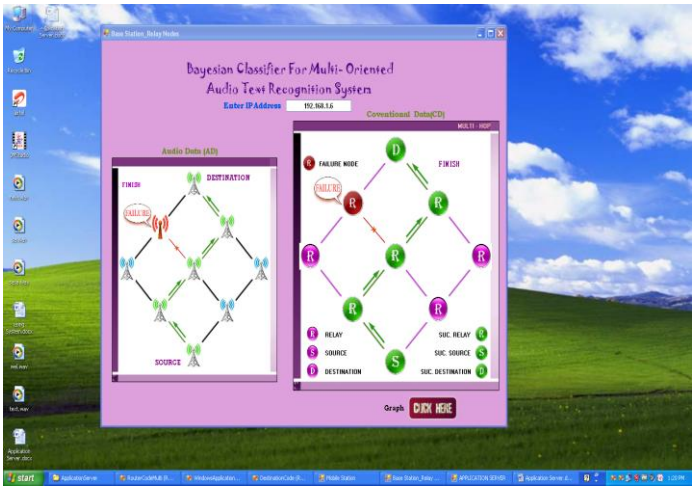

Fig: 6.6 Base Station conversion

CONCLUSION

The growing use of the textual data which needs text to audio mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents Change the thinking from word frequency based vector space to concepts based vector space. The feature selection under concepts, to see if these will help in text to audio categorization. Make the dimensionality reduction more Efficient. Maximum Entropy classifiers use a basic model that is similar to the model used by naive Bayes; however, they employ iterative optimization to find the set of feature weights that maximizes the probability of the training set.

If speech recognition systems someday achieve a generally acceptable level, we may develop for example a communication system where the system may first analyze the speakers' voice and its characteristics, transmit only the character string with some control symbols, and finally synthesize the speech with individual

sounding voice at the other end. Even interpretation from a language to another may became feasible. However, it is obvious that we must wait for several years, maybe decades, until such systems are possible and commonly available.

REFERENCES:

1. Aradhya, V. N. M., Hemantha Kumar, G., & Noushat, S. (2008). Multilingual OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis. Engineering Applications of Artificial Intelligence, 658-668.
2. Chattopadhyay, T., Reddy, V. R., & Garain, U. (2013). Automatic Selection of Binarization Method for Robust OCR. In Proceedings of the ICDAR (pp 1170-1174).
3. Chen, D., & Odobez, J. M. (2005). Video text recognition using sequential Monte Carlo and error voting methods. Pattern Recognition Letters. 1386-1403.
4. Chen, D., Odobez, J. M., & Bourlard, H. (2004). Text detection and recognition in images and video frames. Patten Recognition, 595-608.
5. Crandall, D., Antani, S., & Kasturi, R. (2003). Extraction of special effects caption text events from digital video. International Journal of Document Analysis and Recognition, 38-157.
6. Cunzhao, S., Baihua, X., Chunheng, W., & Yang, Z. (2012). Adaptive Graph Cut Based Binarization of Video Text Images. In Proceedings of the DAS (pp. 58-62).
7. Doermann, D., Liang, J., & Li, H. (2003). Progress in Camera-Based Document Image Analysis. In Proceedings of the ICDAR (pp. 606- 616).

BIBLIOGRAPHY

1. Mrs. Vennila S., M.C.A., M.E., Assistant Professor, Department of Computer Science Auxilium College (Autonomous), Vellore, TamilNadu, India.

2. Ms. Mahalakshmi J., M.Sc Department of Computer Science Auxilium College (Autonomous), Vellore, TamilNadu, India.