NEERUGATTI VISHWANATH, ANIL KUMAR GOGI, P PREM KISHAN, SK. KHAMURUDDEEN, S.V.DEVIKA/International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 2,Mar-Apr 2012, pp.1168-1175

CLASSIFICATION OF SCRIPTS USING VERTICAL STROKE FEATURE

NEERUGATTI VISHWANATH*, ANIL KUMAR GOGI, P PREM KISHAN***, SK. KHAMURUDDEEN****, S.V.DEVIKA*****

*(Assistant professor, Department of ECE, ST. Peters Engineering College, Hyderabad)
** (Assistant professor, Department of ECE, ST. Peters Engineering College, Hyderabad)
*** (Assistant professor, Department of ECE, ST. Peters Engineering College, Hyderabad)
**** (Assistant Professor, Department of ECE, HITAM, Hyderabad, India)
***** (Associate Professor, Department of ECE, HITAM, Hyderabad, India)

ABSTRACT

In a multilingual country like India, a document may contain words of text in more than one language. In this environment, multi lingual Optical Character Recognition (OCR) system is needed to read the documents. It is necessary to identify different layout regions of respective language before feeding the document to the OCR system. This Project Work Proposes prioritized requirements of Andhra Pradesh region. Hence, the documents of Andhra Pradesh Government are generally printed in Telugu, and English languages. Certain documents produced in private and Government sectors, like railways, banks, post-offices of Andhra Pradesh are of tri-lingual (a document having text in three languages) type. When it comes to automation, assuming that there are three OCRs for Telugu, and English languages, a pre-processor is necessary by which the language type of the different texts lines are to be identified.

In this Work, a script identification technique to identify the text lines of Telugu, and English languages from a bilingual document is presented. In this Work, a simple and efficient technique of language identification for Telugu, and English text lines from a printed document is presented. The proposed system is based on the characteristic features of Stroke and Cursive nature of individual text lines of the input document image. The feature extraction is achieved by finding the behavior of Strokes in individual word boundaries from a printed bilingual document image.

***Keyword's*—OCR (Optical Character Recognition), Multilingual, Language, Bilingual language**

I.INTRODUCTION

In recent years, the demand for tools to be able to recognize, search and retrieve written and spoken sources of multilingual information has increased tremendously. With the rapid explosion of online repositories, researchers and developers of cross-lingual search and translation systems can get a lot of resources they need easily from the Internet. However, there are still significant resources that can only be accessed in a printed form, especially for sparse, low density languages. Manipulation and conversion of these printed documents is essential for many researchers and organizations. One of the most important tasks to address with printed documents is the automatic recognition of text, which usually consists of three steps: (1) zone segmentation and text region identification using document layout analysis; (2) text line, word and character segmentation; and (3) optical character recognition (OCR). In the last step, OCR systems are often designed to work on documents with the specific script. In order to parse bilingual or multilingual documents such as patents1 or bilingual dictionaries, or perform multilingual document retrieval, the script must be identified before feeding words to an appropriate OCR system. Language identification is an important topic in pattern recognition and image processing based automatic document analysis and recognition. The objective of Language identification is to translate human identifiable documents to machine identifiable codes.

The world we live in, is getting increasingly interconnected, electronic libraries have become more pervasive and at the same time increasingly automated including the task of presenting a text in any language as automatically translated text in any other language.

Identification of the language in a document image is of primary importance for selection of a specific OCR system processing multi lingual documents .Language identification may seem to be an elementary and simple issue for humans in the real world, but it is difficult for a machine, primarily because different scripts (a script could be a common medium for different languages) are made up of different shaped patterns to produce different character sets. OCR is of special significance for a multi-lingual country like India, where the text portion of the document usually contains information in more than one language. A document containing text information in more than one language is called a multilingual document. For such type of multilingual documents, it is very essential to identify the text language portion of the document, before the analysis of the contents could be made. Although a great number of OCR techniques have been developed over years, almost all existing works on OCR make an important implicit assumption that the language of the document to be processed is known beforehand. Individual OCR tools have been developed to deal best with only one specific language. In an automated environment such document processing systems relying on OCR would clearly need human intervention to select the appropriate OCR package, which is certainly inefficient, undesirable and impractical. A pre-OCR language identification system would enable the correct OCR system to be selected in order to achieve the best character interpretation of the document. This area has not been very widely researched to date, despite its growing importance to the document image processing community and the progression towards the "paperless office". Keeping this drawback in mind, in this paper an attempt has been made to solve a more foundation problem of language identification of a text from a multilingual document, before its contents are automatically read.

Language identification is one of the vision application problems. Generally human system identifies the language in a document using some visible characteristic features such as texture, horizontal lines, vertical lines, which are visually perceivable and appeal to visual sensation. This human visual perception capability has been the motivator for the development of the proposed system. With this context, in this paper, an attempt has been made to simulate the human visual system, to identify the type of the language based on visual clues, without reading the contents of the document. In a multi-lingual country like India (India has 18 regional languages derived from 12 different scripts; a script could be a common medium for different languages), documents like bus reservation forms, passport application forms, examination question papers, bank-challen, language translation books and money-order forms may contain text words in more than one language forms. For such an environment, multi lingual OCR system is needed to read the multilingual documents. To make a multi-lingual OCR system successful, it is necessary to separate portions of different language regions of the document before feeding to individual OCR systems. In this direction, multi lingual document segmentation has strong direct application potential, especially in a multilingual country like India. In the context of Indian languages, some amount of research work has been reported. Further there is a growing demand for automatically processing the documents in every state in India including Andhra Pradesh. Under the three language formula, adopted by most of the Indian states, the document in a state may be printed in its respective official regional language, the national language Hindi and also in English. Accordingly, a document produced in Andhra Pradesh, a state in India, may be printed in its official regional language Telugu, national language Hindi and also in English. For such an environment, multilingual OCR system is needed to read the multilingual documents. To make a multilingual-OCR system successful, it is necessary to develop the multilingual-OCR system that would work in two stages: (i) Identification and separation of different language portions of the document and (ii) Feeding of individual language regions to appropriate OCR system. In this paper, we focus on the first stage of the multilingual-OCR system and present procedures for identification and separation of Telugu and English text portions of the bilingual document produced at Andhra Pradesh, an Indian state.

II. Line and Word Segmentation

Segmentation:

After scanning the document, the document image is subjected to preprocessing for back ground noise elimination, skew correction and binarization to generate the bit map image of the text. The preprocessed image is then segmented into lines, words and characters. The details of line, word and character segmentation are discussed in the following sub-sections.

Line segmentation:

To separate the text lines, from the document image, the horizontal projection profile of

the document image is found. The horizontal projection profile is the histogram of the number of ON pixels along every row of the image. White space between text lines is used to segment the text lines. Figure 1.2 b) shows a sample Telugu document along with its horizontal projection. The projection profile will have valleys of zero height between the text lines. Line segmentation is done at these points.

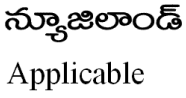

Figure 1.2 a) Input Texts.

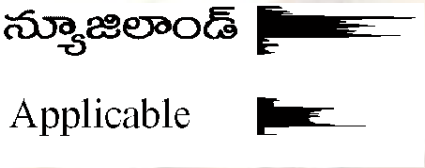

Figure 1.2 b) Horizontal Projection Profiles for line segmentation of Telugu and English texts words.

Word segmentation:

The spacing between the words is used for word segmentation. For Kannada script, spacing between the words is greater than the spacing between characters in a word. The spacing between the words is found by taking the vertical projection profile of an input text line. Vertical projection profile is the sum of ON pixels along every column of the image. A sample input text line and its vertical projection profile is shown in figure1.2 c). From the vertical projection profile it can be observed that, the width of zero-valued valleys is more between the words in the line as compared to the width of zero-valued valleys that exists.

Figure 1.2 c) Vertical Projection Profiles for word segmentation of Telugu and English texts words.

III. CLASSIFICATION OF MULTIPLE SCRIPTS

In India, there are 18 official (Indian constitution accepted) languages. Two or more of these languages may be written in one script. Twelve different scripts are used for writing these languages. Under the three-language formula, many of the Indian documents are written in three languages namely, English, Hindi and the state official language. For example, a money order form in the West-Bengal state is written in English, Hindi and Bangla, because Bangla is the state official language of West-Bengal. Previously we [5] developed a system to identify different scripts from triplets formed by English, Devnagari (Hindi language is written by Devnagari script) and a third state language scripts. A drawback of that system is that we need to know the type of script triplet before using script separation scheme. In this paper we propose a more general scheme to handle all the scripts. If a single document page contains all these twelve scripts, our present method can identify each of them without any prior knowledge of the document. To the best of our knowledge, this is the earliest work of its kind on Indian scripts.

3.1 Identification and Classification of Multi Lingual Script

Every script defines a finite set of text patterns called alphabets. Alphabets of one script are grouped together giving meaningful text information in the form of a word, a text line or a paragraph. Thus, when the alphabets of the same script are combined together to yield meaningful text information, the text portion of the individual script exhibits a distinct visual appearance. The distinct visual appearance of every script is due to the presence of the segments like – horizontal lines, vertical lines, upward curves, downward curves, descendants and so on. The presence of such segments in a particular script is used as visual clues for a human to identify the type of even the unfamiliar script. It was motivated to adopt the idea of human visual perception capability into the proposed model to use the distinct features exhibited by each script. So, the target of this paper is to identify the script type of the texts without reading the contents of the document. By thoroughly observing the structural outline of the characters of the three scripts – Telugu, Devanagari and English, it is observed that the distinct features are present at some specific portion of the characters. It has been found that a distinct characteristic of most of the English characters is the existence of vertical line-like structures [8] and uniform sized characters with each characters having only one component (except "i" and "j" in lower-case).

3.2 Proposed Method Methodology

The script recognition system is composed of four phases.

Preprocessing
Segmentation
Feature extraction
Script classification

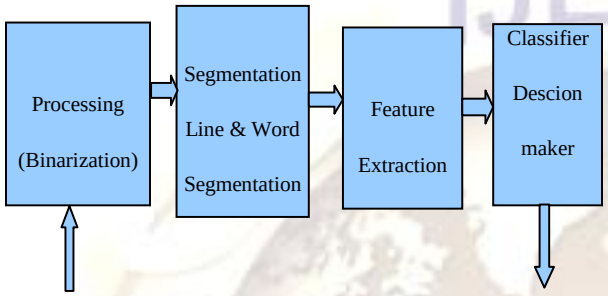

Figure 3.2 a) Block Diagram of proposed Model.

Preprocessing:

The objective of binarization is to automatically choose a threshold that separates the foreground and background information. Selection of a good threshold is often a trial and error process (see figure 3). This becomes particularly difficult in cases where the contrast between text pixels and background is low (for example, text printed on a gray background), when text strokes are very thin resulting in background bleeding into text pixels during digitization, or when the page is not uniformly illuminated during data capture. Many methods have been developed for addressing these problems including those that model the background and foreground pixels as samples drawn from statistical distributions and methods based on spatially varying (adaptive) thresholds. Whether global or adaptive threshold methods are used for binarization, one can seldom expect perfect results. Depending on the quality of the original, there may be gaps in lines, ragged edges on region boundaries, and extraneous pixel regions of ON and OFF values. This fact, that processed results are not perfect, is generally true with other document processing methods, and indeed image processing in general. The recommended procedure is to process as well as possible at each step of processing, but to defer decisions that do not have to be made until later, to avoid making irreparable errors. In later steps, there is more information as a result of processing up to that point, and this provides greater context and higher level descriptions to aid in making correct decisions, and ultimately recognition.

Segmentation

After scanning the document, the document image is subjected to preprocessing for back ground noise elimination, skew correction and binarization to generate the bit map image of the text is necessary but in this project input images created saved as a bit map image. The preprocessed image is then segmented into lines, words and characters. The details of line, word and character segmentation are discussed in the following sub-sections.

Line segmentation

To separate the text lines, from the document image, the horizontal projection profile of the document image is found. The horizontal projection profile is the histogram of the number of ON pixels along every row of the image. White space between text lines is used to segment the text lines. The document image is segmented into several text lines using the valleys of the horizontal projection profiles computed by a row-wise sum of black pixels. The position between two consecutive horizontal projections where the histogram height is least denotes the boundary of a text line.

Word segmentation:

The spacing between the words is used for word segmentation. For Kannada script, spacing between the words is greater than the spacing between characters in a word. The spacing between the words is found by taking the vertical projection profile of an input text line. Vertical projection profile is the sum of ON pixels along every column of the image. From the vertical projection profile it can be observed that, the width of zero-valued valleys is more between the words in the line as compared to the width of zero-valued valleys that exists.

Feature extraction

A feature is a compact representation of the information content in data. In the case of an OCR, shape information of characters are generally characterized as features. The subjective measures like vertical, curvatures, linearity etc. are usually converted as numerical features. Deciding what features to use for a specific script is critical, and it is the phase where most of the development time in an OCR is spent. Below we list some of the popular features employed for Script and recognition. We would like to make it clear that it is not compulsory to work on a small set of features extracted like this.

One could also consider the character image itself as the feature.

The distinct features used in the proposed model are extracted as explained below:

(1) Vertical Stroke: Vertical Stroke is defined as the pixels continuity from top to bottom of script line in straight line like fashion.

(a) Total long vertical strokes: In the proposed model, total long vertical strokes is defined as the number of black pixels from row start to row end in the columns of the Script line.

(b) Total no. of half of the character height vertical strokes: Total no. of half of the character height vertical strokes defined as number of black pixels that have half of the character height.

(2) Cursive nature: Cursive nature is defined as the character set which are in curve shaped.

Feature selection

Feature selection is a process of minimizing the number of features and maximizing the discriminating property of the feature set. Feature selection is a process that aims to identify an optimal subset of relevant features from a large number of features collected in the data set, such that the overall accuracy of classification is increased. In machine learning and statistics, feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models by:

Alleviating the effect of the curse of dimensionality.
Enhancing generalization capability.
Speeding up learning process.
Improving model interpretability.

Feature selection also helps people to acquire better understanding about their data by telling them which are the important features and how they are related with each other. Simple feature selection algorithms are ad hoc, but there are also more methodical approaches. From a theoretical perspective, it can be shown that optimal feature selection for supervised learning problems requires an exhaustive search of all possible subsets of features of the chosen cardinality. If large numbers of features are available, this is impractical. For practical supervised learning algorithms, the search is for a satisfactory set of features instead of an optimal set.

Feature selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset.

Table 3.2 a) Presence and absence of vertical stroke features of Telugu and English text words.
(Yes: means presence and No means absence of that feature. F1: Vertical Stroke; F2: Cursive nature)

Feature / Text words	F1	F2
English	YES	NO
Telugu	NO	YES

.

Script classification

Finally after obtaining the minimum features from the scripts, Decision making rule is used to classify the scripts using vertical stroke and cursive nature features. Decision making: For a given test document image, the features are extracted and the values are computed. In order to classify the given test text line, the input text line is tested for the presence of the features of the script that needs least number of features amongst the three scripts. Accordingly, the test text line is first tested to check whether it is a script because only two features are sufficient for identification. If the test text line is not identified as English script, then the text line is identified as a Telugu script. In this paper, a rule-based classifier is used to classify the test text line into any of the two scripts.

IV. Algorithm

The input document images are created by copying text portion from the internet news papers and hence do not require preprocessing such as noise removal and skew correction. The algorithm for script classification is given below:

Phase I Input Image Processing

1. The Input Image is read as RGB array.
2. The RGB array is converted into Binary Image.

Phase II Line Segmentation

Horizontal profile is calculated and each line in the text document is extracted.

Phase III Word Segmentation

For each line extracted in the phase 2, vertical profile is calculated and each word in the line is extracted.

Phase IV Identification of Vertical Dominance Script and Cursive Dominance Script

1. For each word extracted in the above phase 3 the vertical run length in each column is calculated.

2. The vertical run length for each column is the number of continuous black pixels in the each column.

3. The following are stored for each black run length

(a) The start row value of each black run.

(b) The end row value of each black run.

(c) Whether the black-run length is present in Total half of the character height vertical strokes word.

(d) Whether the start row value of each black run is less than the top line of the word.

(e) Whether the end row value of each black run is more than the bottom line of the word.

(f) If step d and step e is true, then the vertical run length is considered as long run length.

4. The following rules are applied to determine whether the word is English or not.

(a) The number of black runs of the word that are present in the half of the Character height is calculated (Total half width character height).

(b) The number of long lines present in the word is calculated (Total long vertical strokes).

If ((Total no. of long vertical strokes >= Total no. of half width character height *0.5)| (Total half of the character height vertical strokes >= floor (lengthOfLine*0.1)))

It is Vertical Dominance Script. Else

It is Cursive Dominance Script.

Table 1: a) Results showing recognition accuracies:

Input sample	Total English words	Total Telugu words	Total no.of words (A)	Total no. of words recognized correctly (B)	Accuracy =B/A
Sample 1	10	06	16	13	91.4
Sample 2	23	09	32	30	97.3
Sample 3	14	22	36	34	97.3
Sample 4	100	54	154	149	96.7
Sample 5	125	60	165	160	96.9
Sample 6	130	125	255	249	97.6
Sample 7	130	130	260	255	97.6
Sample 8	135	135	270	266	98.5
Sample 9	140	135	275	271	98.5
Sample 10	145	140	285	282	98.9
Sample 11	150	140	290	287	98.9
Sample 12	155	145	300	296	98.6
Sample 13	160	150	310	308	98.6
Sample 14	165	155	320	318	98.6
Sample 15	170	160	330	328	98.6
Sample 16	175	165	340	338	98.6
Sample 17	180	170	350	348	98.6
Sample 18	185	175	360	358	98.6
Sample 19	190	180	370	368	98.6
Sample 20	200	200	400	398	98.6

V.CONCLUSIONS AND FUTURE SCOPE

The proposed model is developed, based on the identification of vertical strokes, which are used to classify text portions of the respective scripts from a printed bilingual document image.Two different set of features are employed successfully for discriminating between Telugu and English words. The first one is total number of long vertical strokes and second is total number of half width character height. The proposed model also includes line and word wise identification models to identify Telugu, and English text words. The major contributions of this work arise from the English and Telugu words. The proposed features are found to be efficient while classifying Telugu and English words. Misclassification amongst Telugu and English text lines is high when the font size of the text line is less than 16 and also when the size of the text line is less than 250X30 pixels. The experimental results show that the two methods are effective and good enough to identify and separate the two language portions of the document, which gives accuracy of 98.4%. This further helps to feed individual language regions to specific OCR system. Our future work is to develop a system that can classify languages based on slant lines or strokes and behavioral aspects to identify individual languages also.

VI.REFERENCES

[1] U.Pal, B.B.Choudhuri, : Script Line Separation From Indian Multi-Script Documents, 5th Int. Conference on Document Analysis and Recognition(IEEE Comput. Soc. Press), 406-409, (1999).

[2] T.N.Tan,: Rotation Invariant Texture Features and their use in Automatic Script Identification, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, pp. 751-756, (1998).

[3] U. Pal, S. Sinha and B. B. Chaudhuri : Multi-Script Line identification from Indian Documents, In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) 0-7695-1960-1/03 © 2003 IEEE, vol.2, pp.880-884, (2003).

[4] Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B. Shet, : Identification of Scripts of Indian Languages by Combining Trainable Classifiers, ICVGIP, Dec.20-22, Bangalore, India, (2000).

[5] S. Chaudhury, R. Sheth, "Trainable script identification strategies for Indian languages", In Proc. 5th Int. Conf. on Document Analysis and Recognition (IEEE Comput. Soc. Press), pp. 657–660, 1999.

[6] Gopal Datt Joshi, Saurabh Garg and Jayanthi Sivaswamy: Script Identification from Indian Documents, LNCS 3872, pp. 255-267, DAS (2006).

[7] S.Basavaraj Patil and N V Subbareddy,: Neural network based system for script identification in Indian documents", Sadhana Vol. 27, Part 1, pp. 83–97. © Printed in India, (2002).

[8] B.V. Dhandra, Mallikarjun Hangarge, Ravindra Hegadi and V.S. Malemath,: Word Level Script Identification in Bilingual Documents through Discriminating Features, IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai, India. Pp.630-635. (2007).

[9] U. Pal and B. B. Chaudhuri, "Automatic separation of Roman, Devnagari and Telugu script lines", Advances in Pattern Recognition and Digital techniques, pp. 447-451, 1999.

[10] Lijun Zhou, Yue Lu and Chew Lim Tan,: Bangla/English Script Identification Based on Analysis of Connected Component Profiles, in proc. 7th DAS, pp. 243-254, (2006).

[11] M. C. Padma and P.Nagabhushan,: Identification and separation of text words of Karnataka, Hindi and English languages through discriminating features, in proc. of Second National Conference on Document Analysis and Recognition, Karnataka, India, pp. 252-260, (2003).

[12] M. C. Padma and P.A.Vijaya,: Language Identification of Kannada, Hindi and English Text Words Through Visual Discriminating Features, International Journal of Computational Intelligence Systems (IJCIS), Volume 1, Issue 2, pp. 116-126, (2008).

[13] Rafael C. Gonzalez, Richard E. Woods and Steven L. Eddins,: Digital Image Processing using MATLAB, Pearson Education, (2004).

[14] Vipin Gupta, G.N. Rathna, K.R. Ramakrishnan,: A Novel Approach to Automatic Identification of Kannada, English and Hindi Words from a Trilingual Document, Int. conf. on Signal and Image Processing, Hubli, pp. 561-566, (2006).

[15] Brunzell H. and Eriksson J., "Feature Reduction for Classification of Multidimensional Data", Pattern Recognition, 33, pp. 1741-1748, 2000.

[16] Sutcliffe, J. P., "On the logical necessity and priority of a monothetic conception of class, and on the consequent inadequacy of polythetic accounts of category and categorization", http://www.db.dk/bh/lifeboat_ko/CONCEPTS/monothetic.html

[17] Shivakumar, Nagabhushan, Hemanthkumar, Manjunath, 2006, "Skew Estimation by Improved Boundary Growing for Text Documents in South Indian Languages", VIVEK- International Journal of Artificial Intelligence, Vol. 16, No. 2, pp 15-21.

[18] Andrew Busch, Wageeh W. Boles and Sridha Sridharan, "Texture for Script Identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 11, pp. 1720-1732, Nov. 2005.

[19] Murali, Vasudev, Hemanthkumar, Nagabhushan, 2006, "Language Independent Skew Detection and Correction of Printed Text Document Images: A Non-rotational Approach", VIVEKInternational Journal of Artificial Intelligence, Vol. 16, No. 2, pp 08-15.

[20] A. L. Spitz, "Determination of script and language content of document images", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 19, No.3, pp. 235–245, 1997. International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009 78

[21] S. L. Wood, X. Yao, K. Krishnamurthy and L. Dang, "Language identification for printed text independent of segmentation", Proc. Int. Conf. on Image Processing, pp. 428–431, 0-8186-7310- 9/95, 1995 IEEE.

[22] J. Hochberg, L. Kerns, P. Kelly and T. Thomas, "Automatic script identification from images using cluster based templates", IEEE Trans. Pattern Anal. Machine Intell. Vol. 19, No. 2, pp. 176–181, 1997.
[23] G. S. Peake and T. N. Tan, "Script and Language Identification from Document Images", Proc. Workshop Document Image Analysis, vol. 1, pp. 10-17, 1997.
[24] A. K. Jain and Y. Zhong, "Page Segmentation using Texture Analysis", Pattern Recognition 29, pp743-770, 1996.

AUTHORS


Neerugatti Viswanath working as Assistant Professor in ST. Peters college of Engineering, Hyderabad and his area of interest is signal processing and communication engineering.


Anil Kumar Gogi working as Assistant Professor in ST. Peters college of Engineering, Hyderabad and his area of interest is signal processing.


Prem Kishan K working as Assistant Professor in ST. Peters college of Engineering, Hyderabad and his area of interest is signal processing.


Mr.SK. Khamuruddeen working as an Assistant Professor in Hyderabad Institute of Technology & Management, His area of interest is VLSI & System Design.


Mrs. S. V. Devika Working as an Associate Professor in Hyderabad Institute of Technology & Management, her area of interest is communications, VLSI & Antenna theory.