Indian Language Document Analysis and Understanding

FOREWORD

Advances in information technology and the wide reach of the internet are radically changing all spheres of activity in our society today. One of the consequences of this is that an increasingly large number of people would be required to interact more frequently with computer systems. To make the man–machine interaction more effective in such situations, it is desirable to have machines capable of handling inputs in a variety of forms such as printed/handwritten paper documents, speech etc. The field of document analysis and understanding is concerned with developing techniques to make it possible for computers to effectively handle (scanned images of) printed documents as input. In spite of widespread use of computers, paper documents will continue to be important for a long (and hopefully transitional) period of time and hence it is important to have computer systems that can seamlessly integrate paper documents with other electronically created ones. There are also other important applications of document image analysis; for example, public digital libraries may require many classical literary works to be processed and made available in digital form.

In a multi-lingual country like India, which has many languages with their own distinctive scripts and rich literary traditions, it is particularly important to develop computer systems that allow users to interact with them in Indian languages. In the context of document image analysis, what we need are techniques for analysing and understanding printed/handwritten documents in Indian languages. Due to the peculiarities of Indian scripts (and languages), solutions that work well for languages such as English would not be applicable, in their totality, for Indian languages. Further, in the Indian context, many documents would contain text of more than one script (for example, English, Hindi and the local language), and hence recognition and segmentation of different scripts from a multilingual document is also an important problem. Thus issues such as recognition of scripts, character recognition in different Indian languages, pre- and post-processing techniques tailored for Indian languages and user-friendly interfaces for better utilisation of the output of document analysis systems, all need attention from Indian scientists working in Image Processing and Pattern Recognition. It is with this motivation that this special issue of *Sādhanā*, depicting the state-of-the-art in Indian Language Document Analysis and Understanding, is planned.

There are eight papers in this special issue. The invited paper authored by Kasturi, O'Gorman and Govindaraju provides a good overview on document image analysis in an authoritative manner outlining all the issues involved. The next two papers deal with complete systems designed for processing printed text documents in a single language. The paper by Chaudhuri, Pal and Mitra, which is also an invited contribution, describes a system for recognition of printed Oriya script. The paper by Ashwin and Sastry describes a system for analysing printed Kannada documents using SVMs for pattern classification. The next paper by Bajaj, Dey and Chaudhury, which is also an invited contribution, describes a method of combining multiple neural network classifiers for robust recognition of handwritten Devanagari numerals.

The next two papers deal with script identification in multilingual documents. The paper by Dhanya, Ramakrishnan and Pati describes a system for identifying the script at each word level in bilingual documents containing Roman and Tamil scripts. The paper by Subba Reddy and Patil describes a neural network based recogniser for identifying the script in a document containing Roman, Devanagari and Kannada scripts.

In most OCR or document analysis applications, the initial recognition accuracy given by the system can be improved through post-processing that makes use of language-specific information. The paper by Lehal and Singh describes such a post-processing system for Gurumukhi.

The final paper in this special issue, authored by Sen and Samudravijaya, is somewhat different in the sense that it does not deal with document image analysis; it deals with one of the possible applications of the results of document image analysis and recognition. This paper describes a text-to-speech system that can read aloud a web document in Hindi or English.

There are research groups in the country working on Indian language document analysis whose work is not reported here. However, we believe that this special issue constitutes a fairly representative sample of the state of the art in this field today. Some of the papers in this issue deal with complete systems for processing printed documents in an Indian language. We hope that such systems will soon reach a stage where they are routinely used in various applications. Further, we hope that the techniques used and reported in the papers in this issue form a good basis for researchers to undertake work on processing of documents in other Indian languages. It will be immensely satisfying if this issue motivates design and development of systems for processing documents in every Indian language in the near future.

February 2002

P S SASTRY
M NARASIMHA MURTY
Guest Editors