Abstract

One of the most common classes of documents which organizations encounter and official communication exploit these days are pre-designed forms. These forms provide a space for entering the transaction details/responses and an understanding of a form therefore includes recognizing the form followed by extracting the information contained in each of its response zones, called variant fields. Since forms are meant for very specific applications, one usually expects very high accuracy and processing speeds. The problem of information processing from forms is usually more structured in nature, and its recognition part is more complex because the texts are usually handwritten or in case of type written they have different fonts, styles and even size. One of the key issues regarding the use of information systems is the acquisition of new information, which often lies in paper documents. In order to provide a suitable solution to this problem, information systems will have to be integrated with paper document processing systems, which are devised to transform printed or handwritten documents into a computer readable form.

Document processing is an important step in the office automation and has emerged as one of the potential research and commercial activity in the field of Computer Sciences. Extensive research has been going on since the past two decades in order to develop a robust document processing system and one of the products of the said activity which has become popular is in the form of an Optical character Recognition System (OCR). Although earlier generally research on paper document processing
has focused on optical character recognition (OCR), in the last decade, it has been widely recognized that text acquisition by means of OCR is only one of the steps of document processing. The process also includes, transformation of the document into an electronic format, the separation of text from graphics, the classification of documents, the identification (or semantic labeling) of some relevant components of the page layout. In the literature, the process of breaking down the bitmap of a scanned paper document (document image) into several layout components is called the document analysis, while the process of attaching semantic (or logic) labels to some layout components is named document understanding. Again, research in the area of document processing for the development of an efficient system in terms of resources and time (as against an OCR) which matches the human system of understanding the objects, forms the thrust area of this research work.

The classification of documents being processed is required for their efficient recognition as it reduces the number of searches and also reduces the chances of error at different stages. Accordingly, in this research a robust classification mechanism for document images based on the layout structure of its different elements which results in cognition based recognition is presented. The document image may contain text only or text as well as tables and images. The classification mechanism carried out in this work is based on a scheme of preserving the structure of the layout of a document image. The algorithms are based on the spatial relationships existing among the visual components present in the document.

The security of the entire information system is presently one of the major concerns which requires a need based solution depending upon the sensitivity of the data involved. The sensitivity of the data/document
varies from organization to organization, and thus the role of personnel with reference to data/document access becomes critical and crucial. People who access it can manipulate it and others can hack. In order to counter the malicious designs of individuals who try to steal (manipulate/hack) information, document/data security can be one of the most challenging issues in an enterprise. In our research the thrust is to ensure that the authorized users are not able to change the data un-authorizably and, accordingly, encryption technique is used to enhance the security of the sensitive data selectively without affecting the normal functioning of the system.

Lastly, it has been found that several independent information systems emerge independently over the period of time resulting in data redundancy as well as replication of services. Information System (IS) planning is one of the most significant activities in a large organization for IS acquisition. The research presents the planning of IS integration plan of the University of Kashmir (UoK) in database aspects. The main objective of this IS integration is to specify techniques in order to link all databases in UoK so that information is interchanged among them seamlessly. As a result of design, all databases share some identical data items, of which information update in batch processing is enough. Thus, database replication and application-program interface (API) are recommended to implement for databases with identical and different database management systems, respectively. In addition, a concept of data warehouse is recommended for implementation to provide summarized information for the UoK which shall assist it in decision making.