Abstract

Document Analysis and Recognition (DAR) targets at automatic extraction of information from documents. *Official document image retrieval* is a crucial step towards office automation. This work considers the processing of office documents for analysis and retrieval. In recent times, vast volumes of documents are being scanned in offices. There is a general need for classification and retrieval of such large-scale digitized documents. In such a scenario, the notion of *Document Image Retrieval* (DIR) is very helpful in providing instant access to such documents. The standard approach for DIR is based only on text. The combined approach for DIR using both visual and textual contents is promising from user's perspective. Segregation of text and non-text contents from document image is of importance for document recognition and retrieval. Official documents can be characterized using textual contents such as dates, keywords as well as using non-textual elements such as logo, stamp, and signature. These non-textual visual patterns uniquely represent legitimacy, identity, and authenticity of documents to a reader. This dissertation addresses issues of text and non-text separation, detection of logo, stamp, signature, and official document image retrieval.

In recent years, most of the official documents are available in color form. This work presents a new color document image dataset called *Scanned Pseudo-Official Data-Set* (SPODS) for detection and retrieval of logo, stamp, and signature. Making the real world office documents available in public domain is difficult due to issues in maintaining confidentiality. However, in the absence of benchmark datasets, it is difficult to evaluate the algorithms developed for the same purpose. Hence, efforts have been made to mimic such documents considering different categories of official records and to build a dataset. The dataset consists of 1088 color document images as representative of regular official communication. It is supplemented by detailed ground truth and metadata for evaluating performance of detection, and retrieval. This is the only document image dataset available in public domain, which provides the ground truth for signature as a set of pixels depicting strokes in the signature.

The data set is augmented by a *Scanned Document Degradation Tool* (SDDT), which is to support degradation of color documents. It helps to simulate some of the document degradation effects such as document skew, blurring, random noise, spot noise, margin noise, and a combination of these degradation processes. The tool is useful for evaluation of document analysis and recognition techniques under degraded conditions. In this work, the performances of text/non-text separation, logo, stamp and signature detection, and document image retrieval are studied using the SPODS dataset and its degraded versions prepared using the SDDT tool. The proposed dataset and tool used for experiments have been made publicly available*.

In this work, novel algorithms for detecting graphics such as logos and stamps in a scanned document image are discussed. A *Spectral Filtering based Text-Graphics Separation algorithm* (SFTGS) is presented here. The property of text

*The dataset and tool are available at `http://www.facweb.iitkgp.ernet.in/~jay/spods/`

that it is the major source of high spatial frequency components in a document image, is exploited in this algorithm to separate text and graphics. The SFTGS algorithm is also extended to support detection of signatures. Next, a *Spectral Filtering based Deep Learning algorithm* (SFDL) for detecting logos and stamps in a scanned document image is presented. In this approach, similar to the SFTGS algorithm, a high frequency filtering is used to suppress text symbols and localizing the candidate regions of interests such as logos and stamps. Finally, these regions are classified using deep convolutional neural network (CNN).

As an initial step of document image retrieval, the thesis presents a novel scheme to separate the text and non-text elements of scanned official documents using new part-based features. It considers the fact that intensity distributions of text and non-text elements in HSV color space are of distinctive nature. A new approach to compute part-based features in S and V channels is proposed. The classification of text and non-text components is performed using majority voting scheme and K-approximate nearest neighbors. Subsequently, the method is extended for detection of logo, stamp, and signature. Experimental results show the effectiveness of the proposed approach.

Finally, an *Official Document Retrieval System* (ODIRS) is presented. The primary task of the ODIRS system is to assist a user in finding the most relevant document images from the database. The system uses text/non-text separation as an initial processing block to simplify document indexing operation. The system targets retrieval of documents using a simple representation in feature space for all types of non-textual queries. It also identifies textual patterns representing dates, keywords for text-based retrieval and supports the formation of combined queries to express complex retrieval requirements. The overall retrieval performance on the SPODS dataset for non-textual attributes in terms of mean average precision (MAP) and mean R-precision (MRP) are 85.6%, and 87.9%, respectively, whereas for degraded SPODS dataset, the overall retrieval performance in terms of MAP and MRP are 77.1%, and 82.2%, respectively. In the case of text based retrieval, the recall for date based document retrieval on SPODS dataset is 97.17%, whereas the recall for keyword based document retrieval is 99.45%. For degraded SPODS dataset, the recall for date based retrieval is 80.34%, whereas the recall for keyword based retrieval is 86.77%.

Keywords: *Dataset; Document degradation; Text/Non-text separation; Part-based features; Spectral filtering; Official document image retrieval*