Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Document Image Analysis

Raval Ajay A.
(H. O. D.)
Shree SPKM BCA & PGDCA College,
Computer Science Dept.,
(Saurashtra University Affiliated)
JETPUR- 360 370.
Mobile: +91-98 251 63674
Email: ravalajay@yahoo.com

Jalwani Ayoob A.
(Lecturer)
BCA / PGDCA Dept.,
Shree G.K. & C.K. Bosamia College,
(Saurashtra University Affiliated)
JETPUR- 360 370.
Mobile: +91-94 275 05192

Karathiya Manoj B.
(H.O.D.)
M.Sc(IT) Dept.,
Shree G.K. & C.K. Bosamia College,
(Saurashtra University Affiliated)
JETPUR- 360 370.
Mobile: +91-94 281 87512

***Abstract.*** **Document image analysis refers to algorithms and techniques that are applied to images of documents to obtain a computer-readable description from pixel data. A well-known document image analysis product is the Optical Character Recognition (OCR) software that recognizes characters in a scanned document. OCR makes it possible for the user to edit or search the document's contents. In this paper we briefly describe various components of a document analysis system. Many of these basic building blocks are found in most document analysis systems, irrespective of the particular domain or language to which they are applied. We hope that this paper will help the reader by providing the background necessary to understand the detailed descriptions of specific techniques presented in other papers in this issue.**

Keywords**. OCR; feature analysis; document processing; graphics recognition; character recognition; layout analysis.**

## I. Introduction

The objective of document image analysis is to recognize the text and graphics components in images of documents, and to extract the intended information as a human would. Two categories of document image analysis can be defined (see figure 1). Textual processing deals with the text components of a document image. Some tasks here are: determining the skew (any tilt at which the document may have been scanned into the computer), finding columns, paragraphs, text lines, and words, and finally recognizing the text (and possibly its attributes such as size, font etc.) by optical character recognition (OCR). Graphics processing deals with the non-textual line and symbol components that make up line diagrams, delimiting straight lines between text sections, company logos etc. Pictures are a third major component of documents, but except for recognizing their location on a page, further analysis of these is usually the task of other image processing and machine vision techniques. After application of these text and graphics analysis techniques, the several megabytes of initial data are culled to yield a much more concise semantic description of the document.

Consider three specific examples of the need for document analysis presented here.

1. Typical documents in today's office are computer-generated, but even so, inevitably by different computers and software such that even their electronic formats are incompatible. Some include formatted text and tables as well as handwritten entries. There are different sizes, from a business card to a large engineering drawing. Document analysis systems recognize types of documents, enable the extraction of their functional parts, and translate from one computer generated format to another.

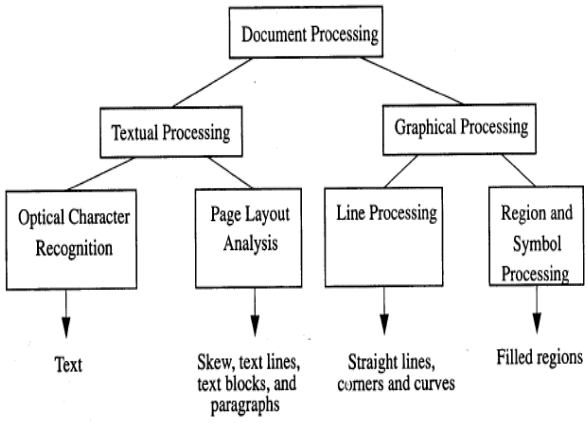


**Figure 1.** A hierarchy of document processing subareas listing the types of document components dealt within each subarea. (Reproduced with permission from O'Gorman & Kasturi 1997.)

2. Automated mail-sorting machines to perform sorting and address recognition have been used for several

 

decades, but there is the need to process more mail, more quickly, and more accurately.

3. In a traditional library, loss of material, misfiling, limited numbers of each copy, and even degradation of materials are common problems, and may be improved by document analysis techniques. All these examples serve as applications ripe for the potential solutions of document image analysis.

Document analysis systems will become increasingly more evident in the form of everyday document systems. For instance, OCR systems will be more widely used to store, search, and excerpt from paper-based documents. Page-layout analysis techniques will recognize a particular form or page format and allow its duplication. Diagrams will be entered from pictures or by hand, and logically edited. Pen-based computers will translate handwritten entries into electronic documents. Archives of paper documents in libraries and engineering companies will be electronically converted for more efficient storage and instant delivery to a home or office computer. Though it will be increasingly the case that documents are produced and reside on a computer, the fact that there are very many different systems and protocols, and also the fact that paper is a very comfortable medium for us to deal with, ensures that paper documents will be with us to some degree for many decades to come. The difference will be that they will finally be integrated into our computerized world.

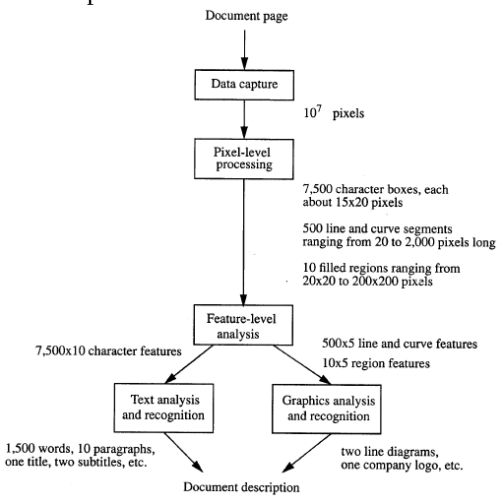


**Figure 2.** A typical sequence of steps for document analysis, along with examples of intermediate and final results and the data size. (Reproduced with permission from O'Gorman & Kasturi 1997.)

Figure 2 illustrates a common sequence of steps in document image analysis. After data capture, the image undergoes pixel-level processing and feature analysis and then text and graphics are treated separately for the recognition of each. We describe these steps briefly in the following sections; the reader is referred to the book, *Document image analysis*, for details (O'Gorman & Kasturi 1997). We conclude this paper by considering the challenges in analysing multilingual documents which is particularly important in the context of Indian language document analysis.

## II. **Data capture**

Data in a paper document are usually captured by optical scanning and stored in a file of picture elements, called pixels that are sampled in a grid pattern throughout the document. These pixels may have values: OFF (0) or ON (1) for binary images, 0–255 for gray-scale images, and 3 channels of 0–255 colour values for colour images. At a typical sampling resolution of 120 pixels per centimeter, a 20 x 30 cm page would yield an image of 2400x3600 pixels. When the document is on a different medium such as microfilm, palm leaves, or fabric, photographic methods are often used to capture images. In any case, it is important to understand that the image of the document contains only raw data that must be further analysed to glean the information.

## III. **Pixel-level processing**

The next step in document analysis is to perform processing on the captured image to prepare it for further analysis. Such processing includes: Thresholding to reduce a grayscale or colour image to a binary image, reduction of noise to reduce extraneous data, segmentation to separate various components in the image, and, finally, thinning or boundary detection to enable easier subsequent detection of pertinent features and objects of interest. After such processing, data are often represented in compact form such as chain-codes and vectors. This pixel-level processing (also called preprocessing and low-level processing in other literature) is the subject of this section. Below are the sub tasks of pixel-level processing…

- ✓ Binarization
- ✓ Noise reduction
- ✓ Segmentation
- ✓ Thinning and region detection
- ✓ Chain coding and vectorization

After pixel-level processing, the raw-image data is converted to a higher level of abstraction; viz., regions representing individual characters, chain codes or vectors representing curve and straight line segments, and boundaries representing large solid objects.

## IV. **Feature-level analysis**

After pixel-level processing has prepared the document image, intermediate features are found from the image to aid in the final step of recognition. At the feature level, thinned and chain-coded data is analysed to detect straight lines, curves, and significant points along the curves. This is a more informative representation that is also closer to how humans would describe the diagram – as lines and curves rather than as ON and OFF points. Curved lines are often approximated by polygonalization. Critical points such as corners and points of high curvature are determined to assist in subsequent analysis for shape recognition. For regions corresponding to individual characters or graphical symbols, local features such as aspect ratio, compactness (ratio of area to square of perimeter), asymmetry, black pixel density, contour smoothness, number of loops, number of line crossings and line ends etc. are computed for input to object recognition stage. Below are the sub tasks of Feature-level analysis…

- ✓ Line and curve fitting

 

- ✓ Polygonalization
- ✓ Critical point detection

## V. **Text document analysis**

There are two main types of analysis that are applied to text in documents. One is optical character recognition (OCR) to derive the meaning of the characters and words from their bit-mapped images, and the other is page-layout analysis to determine the formatting of the text, and from that to derive meaning associated with the positional and functional blocks (titles, subtitles, bodies of text, footnotes etc) in which the text is located. Depending on the arrangement of these text blocks, a page of text may be a title page of a paper, a table of contents of a journal, a business form, or the face of a mail piece. OCR and page layout analysis may be performed separately, or the results from one analysis may be used to aid or correct the other. OCR methods are usually distinguished as being applicable for either machine-printed or handwritten character recognition. Layout analysis techniques are applied to formatted, machine-printed pages, and a type of layout analysis, forms recognition, is applied to machine-printed or handwritten text occurring within delineated blocks on a printed form. In some cases it is necessary to correct the skew of the document which is typically a result of improper paper feeding into the scanner. Skew estimation and layout analysis are discussed briefly in this section. General approaches to OCR are presented in the next section. Below are the sub tasks of Text document analysis…

- ✓ Skew estimation
- ✓ Layout analysis

## VI. **Optical character recognition:**

Optical Character Recognition (OCR) lies at the core of the discipline of pattern recognition where the objective is to interpret a sequence of characters taken from an alphabet. Characters of the alphabet are usually rich in shape. In fact, the characters can be subject to many variations in terms of fonts and handwriting styles. Despite these variations, there is perhaps a basic abstraction of the shapes that identifies any of their instantiations. Developing computer algorithms to identify the characters of the alphabet is the principal task of OCR. The challenge to the research community is the following – while humans can recognize neatly handwritten characters with 100% accuracy, there is no OCR that can match that performance.

OCR difficulty can increase on several counts. Increase in fonts, size of the alphabet set, unconstrained handwriting, touching of adjacent characters, broken strokes due to poor binarization, noise etc. all contribute to the difficulty. figure 3 shows a sample of 0's and 6's that are easily confused by a handwritten digit recognizer.

There are many applications that require the recognition of unconstrained handwriting. A word can be either purely numeric as in the case of a Zip code, or purely alphabetic as in the case of US state abbreviations or mixed as in the number of an apartment (e.g., 1A).

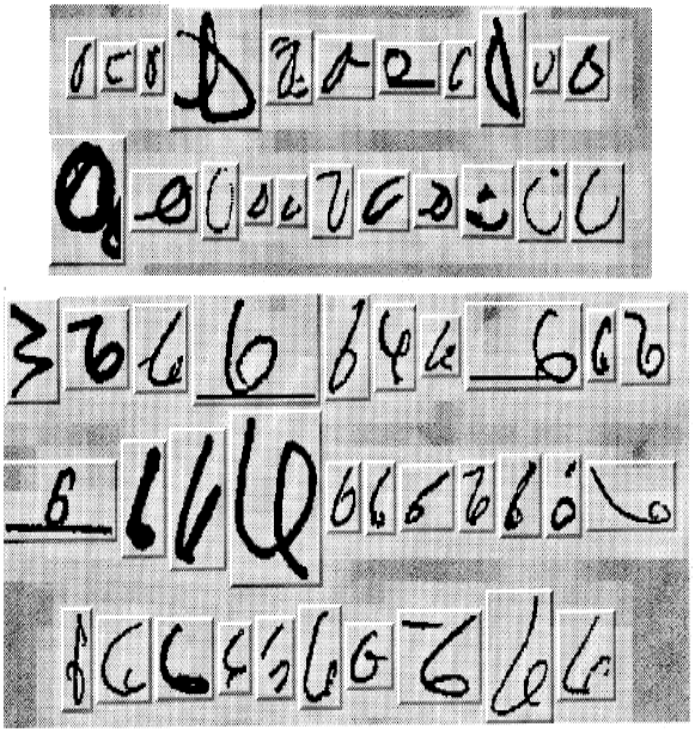

**Figure 3.** Handwritten digits are easily confused

## VII. **Analysis of graphical documents:**

Graphics recognition and interpretation is an important topic in document image analysis since graphics elements pervade textual material, with diagrams illustrating concepts in the text, company logos heading business letters, and lines separating fields in tables and sections of text. The graphics components that we deal with are the binary-valued entities that occur along with text and pictures in documents. We also consider special application domains in which graphical components dominate the document; these include symbols in the forms of lines and regions on engineering diagrams, maps, business charts, fingerprints, musical scores etc. The objective is to obtain information to semantically describe the contents within images of document pages.

Document image analysis can be important when the original document is produced by computer as well. Anyone who has dealt with transport and conversion of computer files knows that compatibility can rarely be taken for granted. Because of the many different languages, proprietary systems, and changing versions of CAD and text formatting packages that are used, incompatibility is especially true in this area. Because the formatted document– that viewed by humans – is semantically the same independent of the language of production, this form is a "protocol-less protocol". If a document system can translate between different machine-drawn formats, the next objective is to translate from hand-drawn graphics. This is analogous to handwriting recognition and text recognition in OCR. When machines can analyse complex hand-drawn diagrams accurately and quickly, the graphics recognition problem will be solved, but there is still much opportunity for research before this goal will be reached.

 

A common sequence of steps taken for document image analysis of graphics interpretation is similar to that for text. Preprocessing, segmentation, and feature extraction methods such as those described in earlier sections are first applied. An initial segmentation step that is generally applied to a mixed text/graphics image is that of text and graphics separation.

An algorithm specifically designed for separating text components in graphics regions irrespective of their orientation is described by Fletcher (1988). This is a Hough transform-based technique that uses the heuristic that text components are collinear. Once text is segmented, typical features extracted from a graphics image include straight lines, curves, and filled regions. After feature extraction, pattern recognition techniques are applied, both structural pattern recognition methods to determine the similarity of an extracted feature to a known feature using geometric and statistical means, and syntactic pattern recognition techniques to accomplish this same task using rules (a grammar) on context and sequence of features.

After this mid-level processing, these features are assembled into entities with some meaning– or semantics – that is dependent upon the domain of the particular application. Techniques used for this include pattern matching, hypothesis and verification, and knowledge-based methods. The semantic interpretation of a graphics element may be different depending on domain; for instance a line may be a road on a map, or an electrical connection of a circuit diagram. Most commercial OCR systems recognize long border and table lines as being different from characters, so no attempt to recognize them as characters is made. Graphics analysis systems for engineering drawings must discriminate between text and graphics (mainly lines). This is usually accomplished very well except for some confusion when characters adjoin lines, causing them to be interpreted as graphics; or when there are small, isolated graphics symbols that are interpreted as characters. Segmentation and analysis of colour-composite multi-layer maps, recognition of the three-dimensional object represented by its orthographic projections in a mechanical part drawing, and construction of a 3-D virtual walk-through from an architectural drawing are some examples of challenges presented to the graphics image analysis researchers. Clearly, much domain-dependent knowledge is applied in essentially all graphics analysis systems.

## VII. **Conclusions**

We presented a brief summary of basic building blocks that comprise a document analysis system. We encourage the reader to refer to cited papers for more detailed descriptions. We hope that this introduction will help the reader by providing the background necessary to understand the contents of the papers that follow in this paper.

References:

- O'Gorman L 1988 Curvilinear feature detection from curvature estimation. *9th Int. Conference on Pattern Recognition*, Rome, Italy, pp 1116–1119
- O'Gorman L 1990 k x k Thinning. *Comput. Vision, Graphics, Image Process.* 51: 195–215
- http://www.itl.nist.gov/iad/894.03/pubs.html

 