Understanding Digital Documents Using Gestalt Properties of Isothetic Components

Shyamosree Pal, Indian Institute of Technology, Kharagpur, India
Partha Bhowmick, Indian Institute of Technology, Kharagpur, India
Arindam Biswas, Bengal Engineering and Science University, Shibpur, Howrah, India
Bhargab B. Bhattacharya, Indian Statistical Institute, Kolkata, India

ABSTRACT

This paper introduces how Gestalt properties can be used for identifying various components in a document image. That the human mind makes a holistic approach to vision rather than a disintegrated approach is shown to be useful for document analysis. Since the major constituent components (textual or non-textual) in a document page are arranged in a rectilinear fashion, rectilinear/isothetic decomposition of different components are made on a document page. After representing the page as a feature set of its polygonal covers corresponding to the distinct regions of interest, each polygon is iteratively decomposed into the sub-polygons tightly enclosing the corresponding sub-components to capture the overall information as well as the necessary details to the desired level of precision. Subsequently, these components and sub-components are analyzed using Gestalt laws/properties, which have been explained in detail in the context of this work. Text regions, tabular structures, and various graphic objects readily admit some of the Gestalt properties. We have tested our algorithm on several benchmark datasets, and some relevant results have been produced here to demonstrate the effectiveness and elegance of the proposed method.

Keywords: Algorithm, Benchmark, Datasets, Document Image, Vision

INTRODUCTION

The theory of Gestalt psychology is based on the idea that an experienced human mind actually makes a holistic approach to vision rather than a disintegrated approach. The mind has the ability to understand an image in such a way that the individual parts of the image produce the collective impression by assuming connections where it does not actually see one but finds necessary to have an overall perception (Sternberg, 2003). Hence, Gestalt psychology
has been used recently in several research paradigms where the visual information has a significant role, e.g., musicology, automated building generation, using semi-autonomous agents to help artists express ideas, designing of web pages, etc. (Leman, 1997; Z. Li, Yan, Ai, & Chen, 2004; Mason, Denzinger, & Carpendale, 2005; Wilson, Russell, Schraefel, & Smith, 2006). We have used Gestalt properties for understanding various digital documents, which is a contemporary problem of the digital era and requires state-of-the-art technologies for its effective solution (Chaudhuri, 2007). Clearly, the solution is largely dependent upon the successful identification of all kinds of structures present in a document image and subsequently finding their associations with different components within a document. Interestingly, a document page has a striking property of admitting a characterization by the rectilinear arrangement of its major constituent components like paragraph, lines, words, tabular structures, graphics, etc. Based on this simple yet useful property, a novel geometric technique is proposed for rectilinear decomposition of different components in a document page, followed by an effective method on indexing and organizing these components for the purpose of efficient retrieval of digital documents. An efficient and meaningful segmentation of the above-mentioned components from a document image is the first step towards indexing of document pages. The second phase involves storing these geometric structures in a scientific way in order to design a robust retrieval system. Given a gray-scale document image, our algorithm performs the segmentation-cum-recognition of its different components by analyzing the geometric features of their respective minimum-area rectilinear/ isothetic polygonal covers corresponding to a few judiciously selected values of the grid spacing, \( g \). As the shape and size of a polygonal cover depends on \( g \) (lower the value of \( g \), tighter is the polygonal cover, and vice versa), and each isothetic polygon is represented by an ordered sequence of its vertices, the spatial relationship of the polygons corresponding to a higher grid spacing with those corresponding to a lower one, is performed using an appropriate geometric analysis of the vertex sequences representing these polygons. Some results on a few datasets are shown in Figures 1 and 2 for a preliminary idea. After discussing the important techniques related with document image segmentation and analysis in the next section, we have explained the major steps of our algorithm in the section of Proposed Method. Experimental results have been presented in the Results section to show the strength and efficacy of the algorithm. Concluding notes and further works that may be benefited out of the proposed algorithm have been pointed out in the section of Conclusion.

**Existing Works on Document Understanding**

In document understanding systems, proper classification of the document into its various constituent parts or zones (text and graphics) is of great importance (Wang, Phillips, & Haralick, 2006). The zone classification technique plays a key role in a document understanding system, which includes text extraction (Xiao & Yan, 2003), OCR (Kopec & Chou, 1994), math recognition (Zanibbi, Blostein, & Cordy, 2002), table understanding (Hu, Kashi, Lopresti, & Wilfong, 2002), image and diagram extraction (Futrelle, Shao, Cieslik, & Grimes, 2003; J. Li, Najmi, & Gray, 2000), etc. An overview of existing works in document image classification in recent times with their performance evaluation may be seen in (Wang et al., 2006). In this section, we review some of the works that are related mainly with our work.

There has been a large number of works devoted to the analysis of digital documents. The analysis has been done on the basis of structure, texture, and frequency-based models. There are works that use various techniques for the segmentation of various structures in a document image. The techniques mainly involve usage of pLSA (probabilistic latent semantic analysis) (Yamaguchi & Maruyama, 2008), Gabor wavelet and kernel-based methods, Voronoi edges, etc. A number of approaches
Barriers to Information Seeking in the Digital Libraries
www.igi-global.com/chapter/barriers-to-information-seeking-in-the-digital-libraries/159603?camid=4v1a