13

Advances in Graphics Recognition

Josep Lladós

13.1 Introduction

Traditionally, the field of document image analysis (DIA) has been defined as a combined sub-discipline of image processing and pattern recognition that proposes theory and practice to the automatic recovery, starting from images of digitized documents, of the syntactic and semantic information that was used to generate them. We can classify documents in terms of three criteria, namely the format, the document contents and the input mode. Considering the format, documents are no longer static, physical entities, but we are moving from paper-based (scanned images of document pages) to electronic documents (e-mails, web pages, DXF, PDF, etc.). Documents can also be classified in terms of the type of information they are conveying. Nagy in his review on document image analysis [1] stated two document categories, namely *mostly text* and *mostly graphics* documents. OCR is at the heart of any system to process mostly-text documents. For those documents, the process also aims at segmenting the layout in paragraphs, columns, lines, words, etc. Examples of mostly-graphics documents include engineering drawings, maps, architectural plans, music scores, schematic diagrams, tables and charts. In such disciplines, graphics are the main way to express information and interact with the machine. It is well known the expression "one picture is worth a thousand words". We can also classify documents in terms of the input mode to create them (*off-line* or *on-line*). On-line documents involve a kind of digital pen interface. In this chapter, we focus on mostly-graphics documents. *Graphics recognition* (GR) can be defined as a branch of document analysis that focuses on the recovery of graphical information in documents. Graphics and text documents are not disjoint categories. Although being considered graphic documents, most of them also can contain textual items, so OCR-related processes and graphics

recognition ones sometimes are collaborative tasks. *Graphics* or *graphical information* is a broad concept. In a logical level, graphics combine primitives that generally are lines, regions or simple shapes. In a functional or semantic level, graphical information consists of a set of compound objects that, in terms of domain-dependent knowledge, have a particular meaning in the context where they appear.

Graphics recognition, as document analysis, is essentially an engineering discipline. Therefore, from a more general point of view, it can be seen as a component in the document engineering lifecycle. Because of that, graphics recognition systems are usually problem or application oriented. Thus a wide variety of pattern recognition and image processing techniques are used as components in graphics recognition systems. Figure 13.1 is an attempt to summarize the major concepts involved in the graphics recognition domain in a coherent component chart. From the methodological point of view, three levels are involved in the processing of a graphical document, namely the *early*, the *structure* and the *semantic* processing, respectively.

Early-level processing can be organized in two sub-categories. First, *image filtering*, i.e. pixel-based processes for noise removal, binarization, or image enhancement. In general, the above pixel-oriented methods are not exclusive to the graphics recognition domain but common among all the document image Analysis areas. A good overview of such techniques can be seen in [2]. The second sub-category consists of the set of tasks devoted to *primitive segmentation*. In the GR domain, a primitive is considered to be a straight segment, an arc, a loop, or a solid region, i.e. graphical tokens that are combined to form graphical entities. We can notice that most of graphic documents consist in line-drawing structures. Because of that, primitive extraction in GR can usually involve a vectorization process. Other primitives than straight segments are also considered in some GR systems. Arc detection is a particular topic of interest [3,4]. Solid regions are another frequent type of primitives. Typical examples are notes in musical score recognition, text or small symbols in map-to-GIS conversion systems. Different criteria such as connected components, colour, area, etc. can be used to segment them. In other cases, the segmentation is also guided by domain-dependent rules as for example when notes are separated from staff lines in musical scores.

As Dori also discussed [5], GR systems have two additional levels directly related to structure (how graphical entities are constructed) and function (what they do or what they mean in the context where they appear). In other words, a syntactic level where the recognition involves locating the graphic entities in the document and classifying them into the various classes by their shape and context, and a semantic level where, using domain-dependent information, a meaning is assigned to the syntactically recognized entities. Syntactic processing includes several tasks related to the classification of graphical entities. Symbol recognition is the main activity. Signatures compactly represent symbol classes mainly for indexing