Part C

Text Recognition

Text recognition can be thought of as the patriarch of document image analysis. The earliest patents and reading machines were all focused on reading individual characters, primarily machine-printed characters in the English language. Thus it is no surprise that this part of the handbook is the largest, with seven chapters. It also shows how the field has drastically changed over the past four decades in individual chapters on essential sub-problems such as segmentation and language identification, while problems such as word recognition, handwriting recognition, and recognition of foreign languages have also evolved to the point where they are considered their own disciplines in the field.

For readers that are interested in any of these topics, the chapters in this part should not necessarily be seen as independent. There is a great deal of overlapping and the techniques that are prominent in machine print recognition, for example, may in fact be very relevant to related issues in Asian or Middle Eastern recognition. Readers are encouraged to at least skim each chapter to get a better picture of issues before delving deeper into a particular chapter.

One all-encompassing chapter is that of text segmentation contributed by Nicola Nobile and Ching Suen (▶Chap. 8 (Text Segmentation for Document Recognition)). It has been widely claimed by commercial OCR vendors that over 80 % of the machine print errors that occur in OCR systems are due to segmentation. Noise tends to cause characters to touch or break, and lines to overlap. The chapter addresses these basic problems, but then considers problems unique to historic and handwritten material.

Along with broadening the field beyond machine print recognition, document image analysis has expanded to deal with many languages, and writing scripts and the plethora of fonts brought on by advanced desktop publishing. ▶Chapter 9 (Language, Script and Font Recognition) was contributed by Umapada Pal. It highlights the challenges and solutions provided by current approaches, and provides a framework for continuing to address their problems.

▶Chapters 10 (Machine-Printed Character Recognition), ▶11 (Handprinted Character and Word Recognition), and ▶12 (Continuous Handwritten Script Recognition) focus on the core of text recognition. ▶Chapter 10 (Machine-Printed

Character Recognition) contributed by Premkuram Natarajan provides historical perspective on machine-printed character recognition. This was one of the first areas to be "commercialized" and while it is seen as a solved problem under ideal circumstances, research is still active on the more challenging aspects of noisy and poor quality data. Venu Govindaraju and Sergey Tulyakov address the complementary problem of handprinted text in ▶Chap. 11 (Handprinted Character and Word Recognition). The problem is often seen as one step harder than machine print, but as you will read in this chapter, the techniques required are fundamentally different on many levels. ▶Chapter 12 (Continuous Handwritten Script Recognition) authored by Horst Bunke and Volkmar Frinken focuses on a problem more commonly referred to as cursive script. The problem introduced fundamental challenges in dealing with different writing styles and the inability to perform accurate segmentation. Readers will appreciate the challenges of automating humans' advanced ability to recognize content even when a majority of the individual characters are illegible.

Finally, Part C concludes with two chapters that focus on specific regions of the world, in part because of the fundamental differences in the scripts and character sets in both machine and handwritten material. In ▶Chap. 13 (Middle Eastern Character Recognition), Abdel Belaïd and Mohamed Imran Razzak address problems arising with the connected scripts of the Middle East. Even for machine print, the connectedness of characters and the variations that occur, based on linguistic context, have forced the community to expand the thinking when it comes to both segmentation and classification. ▶Chapter 14 (Asian Character Recognition) is authored by Ranga Setlur and Zhixin Shi. In addition to segmentation differences with other languages, languages such as Chinese have characters derived from ideograms, and typically have a character set that are orders of magnitude larger than other languages. This provides a unique set of challenges for classification that must be addressed by systems moving forward.