Optical Character Recognition

András Kornai

Final version

The recognition of handwritten or printed text by computer is referred to as O[ptical] C[haracter] R[ecognition]. When the input device is a digitizer tablet that transmits the signal in real time (as in pen-based computers and personal digital assistants) or includes timing information together with pen position (as in signature capture) we speak of *dynamic* recognition. When the input device is a still camera or a scanner, which captures the position of digital ink on the page but not the order in which it was laid down, we speak of *static* or *image-based* OCR.

Dynamic OCR is an increasingly important modality in H[uman] C[omputer] I[interaction], and the difficulties encountered in the process are largely similar to those found in other HCI modalities, in particular → *Speech Recognition.* The stream of position/pen pressure values output by the digitizer tablet is analogous to the stream of speech signal vectors output by the audio processing front end, and the same kinds of lossy data compression techniques, including cepstral analysis, linear predictive coding, and vector quantization, are widely employed for both.

Static OCR encompasses a range of problems that have no counterpart in the recognition of spoken or signed language, usually collected under the heading of *page decomposition* or *layout analysis.* These include both the separation of linguistic material from photos, line drawings, and other non-linguistic information, establishing the local horizontal and vertical axes (deskewing), and the appropriate grouping of titles, headers, footers, and other material set in a font different from the main body of the text. Another OCR-specific problem is that we often find different scripts, such as Kanji and Kana, or Cyrillic and Latin, in the same running text.

While the early experimental OCR systems were often rule-based, by the eighties these have been completely replaced by systems based on statistical → *Pattern Recognition.* For clearly segmented printed materials such techniques offer virtually error-free OCR for the most important alphabetic systems including variants of the Latin, Greek, Cyrillic, and Hebrew alphabets. However, when the number of symbols is large, as in the Chinese or Korean writing systems, or the symbols are not separated from one another, as in Arabic or Devanagari print, OCR systems are still far from the error rates of human readers, and the gap between the two is also evident when the quality of the image is compromised e.g. by fax transmission. Until these problems are resolved, OCR can not play the pivotal role in the transmission of cultural heritage to the digital age that it is often assumed to have.

In the recognition of handprint, algorithms with succesive segmentation, classification, and identification (language modeling) stages are still in the lead. For cursive handwriting, → *Hidden Markov Models* that make the segmentation, classification, and identification decisions in parallel have proven superior, but performance still leaves much to be desired, both because the spatial and the temporal aspects of the writen signal are not necessarily in lockstep (*discontinuous constituents* arising e.g. at the crossing of t-s and dotting of i-s) and because the inherent variability of handwriting is far greater than that of speech, to the extent that we often see illegible handwriting but rarerly hear unintelligeble speech.

For cursive machine-print see e.g. Bazzi et al 1999. The state of the art in handwriting recognition is closely tracked by the International Workshop on Frontiers of Handwriting Recognition (IWFHR). For language modeling in OCR see Kornai 1994. A good general introduction to the problems of page decomposition is O'Gorman and Kasturi (1995), and to OCR in general Bunke and Wang (1997).

References Bazzi, Issam, Richard Schwartz and John Makhoul (1999) An omnifont open-vocabulary OCR system for English and Arabic. Pattern Analysis and Machine Intelligence 21 495-504

Bunke, Horst, and Patrick S.P. Wang (1997) Handbook of Character Recognition and Document Image Analysis. World Scientific, Singapore

Kornai, András (1994) Language models: where are the bottlenecks? AISB Quarterly 88 36-40

O'Gorman, Lawrence and Rangachar Kasturi (1995) Document Image Analysis IEEE Computer Society Press, Los Alamitos CA

Index Optical Character Recognition (OCR), dynamic OCR, image-based OCR, digitizer tablet, pen-based computers, personal digital assistant (PDA), signal vector, language modeling, data compression, page decomposition, discontinuous constituent, variability, layout analysis, Human-Computer Interaction (HCI), Arabic script, Latin script, Hebrew script, Greek script, Cyrillic script, Devanagari script, Chinese script (Han writing), Korean script (Hangul), Kanji, Kana, font