Document Image Analysis and Digital Libraries

Henry S. Baird
Computer Science & Engineering Dept
Lehigh University
Bethlehem, PA 18015 USA

baird@parc.xerox.com

Abstract

The rapid growth of digital libraries (DLs) worldwide poses many new challenges for document image analysis (DIA) research and development. DLs promise to offer more people access to larger document collections, and at far greater speed, than physical libraries can. But DLs also tend, for many reasons, to serve poorly, or even to omit entirely, many types of non-digital human--legible media such as originally printed and handwritten documents. These documents, in their original physical (undigitized) form, are readily - if not always quickly - legible, searchable, and browseable, whereas in the form of images accessed through DLs they often lose many of their original advantages while of course lacking many advantages of symbolically encoded information. This talk explores these issues and illustrates them with brief case studies arising from his experience as a DIA researcher in collaboration with several DL projects in the US. Difficult open DIA technical problems in DL applications are identified in, e.g., the contrasting advantages of paper and digital displays, during image capture, early processing, recognition, analysis, presentation, and retrieval - and in personal and interactive applications.

These support the conclusion that the international image understanding R&D community is urgently needed (because uniquely qualified) to provide new technology to help rescue from neglect - even, in many cases, eventual oblivion - the world's vast culturally irreplaceable legacy paper document collections.