Simone Marinai
Andreas Dengel (Eds.)

LNCS 3163

Document Analysis Systems VI

6th International Workshop, DAS 2004
Florence, Italy, September 2004
Proceedings

Springer

Simone Marinai Andreas Dengel (Eds.)

Document Analysis Systems VI

6th International Workshop, DAS 2004
Florence, Italy, September 8 - 10, 2004
Proceedings

Springer

Volume Editors

Simone Marinai
Università di Firenze, Dipartimento di Sistemi e Informatica
Via S. Marta, 3 - 50139 Firenze, Italy
E-mail: marinai@dsi.unifi.it

Andreas Dengel
German Research Center for Artificial Intelligence (DFKI)
P.O.Box 2080, 67608 Kaiserslautern, Germany
E-mail: Andreas.Dengel@dfki.de

Library of Congress Control Number: 2004111168

CR Subject Classification (1998): I.5, H.3, I.4, I.7, J.1, J.2

ISSN 0302-9743
ISBN 3-540-23060-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11321026 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3163

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Lecture Notes in Computer Science

For information about Vols. 1–3089

please contact your bookseller or Springer

Vol. 3220: J.C. Lester, R.M. Vicari, F. Paraguaçu (Eds.), Intelligent Tutoring Systems. XXI, 920 pages. 2004.

Vol. 3208: H.J. Ohlbach, S. Schaffert (Eds.), Principles and Practice of Semantic Web Reasoning. VII, 165 pages. 2004.

Vol. 3207: L.T. Jang, M. Guo, G.R. Gao, N.K. Jha, Embedded and Ubiquitous Computing. XX, 1116 pages. 2004.

Vol. 3205: N. Davies, E. Mynatt, I. Siio (Eds.), UbiComp 2004: Ubiquitous Computing. XVI, 452 pages. 2004.

Vol. 3203: J. Becker, M. Platzner, S. Vernalde (Eds.), Field Programmable Logic and Application. XXX, 1198 pages. 2004.

Vol. 3198: G.-J. de Vreede, L.A. Guerrero, G. Marín Raventós (Eds.), Groupware: Design, Implementation and Use. XI, 378 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004. (Subseries LNAI).

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004. (Subseries LNAI).

Vol. 3186: Z. Bellahsène, T. Milo, M. Rys, D. Suciu, R. Unland (Eds.), Database and XML Technologies. X, 235 pages. 2004.

Vol. 3184: S. Katsikas, J. Lopez, G. Pernul (Eds.), Trust and Privacy in Digital Business. XI, 299 pages. 2004.

Vol. 3183: R. Traunmüller (Ed.), Electronic Government. XIX, 583 pages. 2004.

Vol. 3182: K. Bauknecht, M. Bichler, B. Pröll (Eds.), E-Commerce and Web Technologies. XI, 370 pages. 2004.

Vol. 3181: Y. Kambayashi, M. Mohania, W. W\"oß (Eds.), Data Warehousing and Knowledge Discovery. XIV, 412 pages. 2004.

Vol. 3180: F. Galindo, M. Takizawa, R. Traunmüller (Eds.), Database and Expert Systems Applications. XXI, 945 pages. 2004.

Vol. 3178: W. Jonker, M. Petkovic (Eds.), Secure Data Management. VIII, 219 pages. 2004.

Vol. 3177: Z.R. Yang, H. Yin, R. Everson (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2004. XVIII, 852 pages. 2004.

Vol. 3175: C.E. Rasmussen, H.H. Bülthoff, B. Schölkopf, M.A. Giese (Eds.), Pattern Recognition. XVIII, 581 pages. 2004.

Vol. 3174: F. Yin, J. Wang, C. Guo (Eds.), Advances in Neural Networks - ISNN 2004. XXXV, 1021 pages. 2004.

Vol. 3172: M. Dorigo, M. Birattari, C. Blum, L. M. Gambardella, F. Mondada, T. Stützle (Eds.), Ant Colony, Optimization and Swarm Intelligence. XII, 434 pages. 2004.

Vol. 3170: P. Gardner, N. Yoshida (Eds.), CONCUR 2004 - Concurrency Theory. XIII, 529 pages. 2004.

Vol. 3166: M. Rauterberg (Ed.), Entertainment Computing – ICEC 2004. XXIII, 617 pages. 2004.

Vol. 3163: S. Marinai, A. Dengel (Eds.), Document Analysis Systems VI. XI, 564 pages. 2004.

Vol. 3159: U. Visser, Intelligent Information Integration for the Semantic Web. XIV, 150 pages. 2004. (Subseries LNAI).

Vol. 3158: I. Nikolaidis, M. Barbeau, E. Kranakis (Eds.), Ad-Hoc, Mobile, and Wireless Networks. IX, 344 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004. (Subseries LNAI).

Vol. 3156: M. Joye, J.-J. Quisquater (Eds.), Cryptographic Hardware and Embedded Systems - CHES 2004. XIII, 455 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004. (Subseries LNAI).

Vol. 3154: R.L. Nord (Ed.), Software Product Lines. XIV, 334 pages. 2004.

Vol. 3153: J. Fiala, V. Koubek, J. Kratochvíl (Eds.), Mathematical Foundations of Computer Science 2004. XIV, 902 pages. 2004.

Vol. 3152: M. Franklin (Ed.), Advances in Cryptology – CRYPTO 2004. XI, 579 pages. 2004.

Vol. 3150: G.-Z. Yang, T. Jiang (Eds.), Medical Imaging and Augmented Reality. XII, 378 pages. 2004.

Vol. 3149: M. Danelutto, M. Vanneschi, D. Laforenza (Eds.), Euro-Par 2004 Parallel Processing. XXXIV, 1081 pages. 2004.

Vol. 3148: R. Giacobazzi (Ed.), Static Analysis. XI, 393 pages. 2004.

Vol. 3146: P. Érdi, A. Esposito, M. Marinaro, S. Scarpetta (Eds.), Computational Neuroscience: Cortical Dynamics. XI, 161 pages. 2004.

Vol. 3144: M. Papatriantafilou, P. Hunel (Eds.), Principles of Distributed Systems. XI, 246 pages. 2004.

Vol. 3143: W. Liu, Y. Shi, Q. Li (Eds.), Advances in Web-Based Learning – ICWL 2004. XIV, 459 pages. 2004.

Vol. 3142: J. Diaz, J. Karhumäki, A. Lepistö, D. Sannella (Eds.), Automata, Languages and Programming. XIX, 1253 pages. 2004.

Vol. 3140: N. Koch, P. Fraternali, M. Wirsing (Eds.), Web Engineering. XXI, 623 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004. (Subseries LNAI).

Vol. 3138: A. Fred, T. Caelli, R.P.W. Duin, A. Campilho, D.d. Ridder (Eds.), Structural, Syntactic, and Statistical Pattern Recognition. XXII, 1168 pages. 2004.

Vol. 3137: P. De Bra, W. Nejdl (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems. XIV, 442 pages. 2004.

Vol. 3136: F. Meziane, E. Métais (Eds.), Natural Language Processing and Information Systems. XII, 436 pages. 2004.

Vol. 3134: C. Zannier, H. Erdogmus, L. Lindstrom (Eds.), Extreme Programming and Agile Methods - XP/Agile Universe 2004. XIV, 233 pages. 2004.

Vol. 3133: A.D. Pimentel, S. Vassiliadis (Eds.), Computer Systems: Architectures, Modeling, and Simulation. XIII, 562 pages. 2004.

Vol. 3132: B. Demoen, V. Lifschitz (Eds.), Logic Programming. XII, 480 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004. (Subseries LNAI).

Vol. 3130: A. Syropoulos, K. Berry, Y. Haralambous, B. Hughes, S. Peter, J. Plaice (Eds.), TeX, XML, and Digital Typography. VIII, 265 pages. 2004.

Vol. 3129: Q. Li, G. Wang, L. Feng (Eds.), Advances in Web-Age Information Management. XVII, 753 pages. 2004.

Vol. 3128: D. Asonov (Ed.), Querying Databases Privately. IX, 115 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004. (Subseries LNAI).

Vol. 3126: P. Dini, P. Lorenz, J.N.d. Souza (Eds.), Service Assurance with Partial and Intermittent Resources. XI, 312 pages. 2004.

Vol. 3125: D. Kozen (Ed.), Mathematics of Program Construction. X, 401 pages. 2004.

Vol. 3124: J.N. de Souza, P. Dini, P. Lorenz (Eds.), Telecommunications and Networking - ICT 2004. XXVI, 1390 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004. (Subseries LNAI).

Vol. 3122: K. Jansen, S. Khanna, J.D.P. Rolim, D. Ron (Eds.), Approximation, Randomization, and Combinatorial Optimization. IX, 428 pages. 2004.

Vol. 3121: S. Nikoletseas, J.D.P. Rolim (Eds.), Algorithmic Aspects of Wireless Sensor Networks. X, 201 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004. (Subseries LNAI).

Vol. 3118: K. Miesenberger, J. Klaus, W. Zagler, D. Burger (Eds.), Computer Helping People with Special Needs. XXIII, 1191 pages. 2004.

Vol. 3116: C. Rattray, S. Maharaj, C. Shankland (Eds.), Algebraic Methodology and Software Technology. XI, 569 pages. 2004.

Vol. 3114: R. Alur, D.A. Peled (Eds.), Computer Aided Verification. XII, 536 pages. 2004.

Vol. 3113: J. Karhumäki, H. Maurer, G. Paun, G. Rozenberg (Eds.), Theory Is Forever. X, 283 pages. 2004.

Vol. 3112: H. Williams, L. MacKinnon (Eds.), Key Technologies for Data Management. XII, 265 pages. 2004.

Vol. 3111: T. Hagerup, J. Katajainen (Eds.), Algorithm Theory - SWAT 2004. XI, 506 pages. 2004.

Vol. 3110: A. Juels (Ed.), Financial Cryptography. XI, 281 pages. 2004.

Vol. 3109: S.C. Sahinalp, S. Muthukrishnan, U. Dogrusoz (Eds.), Combinatorial Pattern Matching. XII, 486 pages. 2004.

Vol. 3108: H. Wang, J. Pieprzyk, V. Varadharajan (Eds.), Information Security and Privacy. XII, 494 pages. 2004.

Vol. 3107: J. Bosch, C. Krueger (Eds.), Software Reuse: Methods, Techniques and Tools. XI, 339 pages. 2004.

Vol. 3106: K.-Y. Chwa, J.I. Munro (Eds.), Computing and Combinatorics. XIII, 474 pages. 2004.

Vol. 3105: S. Göbel, U. Spierling, A. Hoffmann, I. Iurgel, O. Schneider, J. Dechau, A. Feix (Eds.), Technologies for Interactive Digital Storytelling and Entertainment. XVI, 304 pages. 2004.

Vol. 3104: R. Kralovic, O. Sykora (Eds.), Structural Information and Communication Complexity. X, 303 pages. 2004.

Vol. 3103: K. Deb, e. al. (Eds.), Genetic and Evolutionary Computation – GECCO 2004. XLIX, 1439 pages. 2004.

Vol. 3102: K. Deb, e. al. (Eds.), Genetic and Evolutionary Computation – GECCO 2004. L, 1445 pages. 2004.

Vol. 3101: M. Masoodian, S. Jones, B. Rogers (Eds.), Computer Human Interaction. XIV, 694 pages. 2004.

Vol. 3100: J.F. Peters, A. Skowron, J.W. Grzymała-Busse, B. Kostek, R.W. Świniarski, M.S. Szczuka (Eds.), Transactions on Rough Sets I. X, 405 pages. 2004.

Vol. 3099: J. Cortadella, W. Reisig (Eds.), Applications and Theory of Petri Nets 2004. XI, 505 pages. 2004.

Vol. 3098: J. Desel, W. Reisig, G. Rozenberg (Eds.), Lectures on Concurrency and Petri Nets. VIII, 849 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004. (Subseries LNAI).

Vol. 3096: G. Melnik, H. Holz (Eds.), Advances in Learning Software Organizations. X, 173 pages. 2004.

Vol. 3095: C. Bussler, D. Fensel, M.E. Orlowska, J. Yang (Eds.), Web Services, E-Business, and the Semantic Web. X, 147 pages. 2004.

Vol. 3094: A. Nürnberger, M. Detyniecki (Eds.), Adaptive Multimedia Retrieval. VIII, 229 pages. 2004.

Vol. 3093: S. Katsikas, S. Gritzalis, J. Lopez (Eds.), Public Key Infrastructure. XIII, 380 pages. 2004.

Vol. 3092: J. Eckstein, H. Baumeister (Eds.), Extreme Programming and Agile Processes in Software Engineering. XVI, 358 pages. 2004.

Vol. 3091: V. van Oostrom (Ed.), Rewriting Techniques and Applications. X, 313 pages. 2004.

Preface

This volume contains papers selected for presentation at the 6th IAPR Workshop on Document Analysis Systems (DAS 2004) held during September 8–10, 2004 at the University of Florence, Italy. Several papers represent the state of the art in a broad range of "traditional" topics such as layout analysis, applications to graphics recognition, and handwritten documents. Other contributions address the description of complete working systems, which is one of the strengths of this workshop. Some papers extend the application domains to other media, like the processing of Internet documents.

The peculiarity of this 6th workshop was the large number of papers related to digital libraries and to the processing of historical documents, a taste which frequently requires the analysis of color documents. A total of 17 papers are associated with these topics, whereas two years ago (in DAS 2002) only a couple of papers dealt with these problems.

In our view there are three main reasons for this new wave in the DAS community. From the scientific point of view, several research fields reached a thorough knowledge of techniques and problems that can be effectively solved, and this expertise can now be applied to new domains. Another incentive has been provided by several research projects funded by the EC and the NSF on topics related to digital libraries. Last but not least, the organization of focused events, like the recent DIAL workshop chaired by Henry Baird and Venu Govindraju in Palo Alto (CA), had a strong impact on the definition of new research directions. However, it is indeed a lucky coincidence that this new trend in DAS research emerged in this edition organized in a town such as Florence, which keeps such an exceptional artistic and cultural heritage.

We received a total of 79 submissions from 19 countries, and we selected 31 oral presentations and 22 posters highlighted with short oral introductions. As a supplement to this proceedings, notes from the workshop discussions and other material related to presented papers will be posted on the DAS 2004 website: http://www.dsi.unifi.it/DAS04. Each paper was reviewed by three reviewers whom we would like to warmly thank here. We should mention the valuable support and hints provided by members of the Program Committee and past DAS chairs. We also wish to acknowledge the generosity of our sponsors: the International Association for Pattern Recognition, the University of Florence, the DFKI, ABBYY, Hitachi, and Siemens.

Special thanks are due to Alessio Ceroni, Cristina Dolfi, and Emanuele Marino for their invaluable contributions to the local organization.

June 2004

Simone Marinai
Andreas Dengel

Organization

Workshop Co-chairs

Simone Marinai	University of Florence, Italy
Andreas Dengel	DFKI, Germany

Program Committee

Apostolos Antonacopoulos	University of Liverpool, UK
Henry Baird	Lehigh University, USA
Francesca Cesarini	University of Florence, Italy
David Doermann	University of Maryland, USA
Andrew Downton	University of Essex, UK
Hiromichi Fujisawa	Hitachi Central Research Laboratory, Japan
Jianying Hu	IBM T.J. Watson Research Center, USA
Rolf Ingold	University of Fribourg, Switzerland
Ramanujan Kashi	Avaya Labs Research, USA
Koichi Kise	Osaka Prefecture University, Japan
Dan Lopresti	Lehigh University, USA
Donato Malerba	University of Bari, Italy
Udo Miletzki	Siemens Dematic, Germany
Yasuaki Nakano	Kyushu University, Japan
Lambert Schomaker	Rijksuniversiteit Groningen, The Netherlands
Giovanni Soda	University of Florence, Italy
Larry Spitz	Document Recognition Technologies, New Zealand
Karl Tombre	LORIA-INPL, France
Luc Vincent	LizardTech, USA
Marcel Worring	University of Amsterdam, The Netherlands

Additional Referees

Annalisa Appice	Dimosthenis Karatzas	T.R. Roth-Berghofer
Margherita Berardi	Michele Lapi	Jane Snowdon
Alain Biem	Larry O'Gorman	Salvatore Tabbone
Thomas Breuel	Huanfeng Ma	Yefeng Zheng
Michelangelo Ceci	Gérald Masini	Gary Zi
Philippe Dosch	Eugene Ratzlaff	
Stefan Jaeger	Maurizio Rigamonti	

Table of Contents

Digital Libraries

Historical Documents

Layout Analysis

Color Documents

Handwritten Documents

Graphics Recognition

Internet Documents

Document Analysis Systems

Applications

Document Analysis Systems for Digital Libraries: Challenges and Opportunities

Henry S. Baird[1], Venugopal Govindaraju[2], and Daniel P. Lopresti[1]

[1] CSE Department, Lehigh University, Bethlehem, PA, USA
{baird,lopresti}@cse.lehigh.edu
[2] CEDAR, University at Buffalo, SUNY, Buffalo, NY, USA
govind@cedar.buffalo.edu

Abstract. Implications of technical demands made within digital libraries (DL's) for document image analysis systems are discussed. The state-of-the-art is summarized, including a digest of themes that emerged during the recent International Workshop on Document Image Analysis for Libraries. We attempt to specify, in considerable detail, the essential features of document analysis systems that can assist in: (a) the creation of DL's; (b) automatic indexing and retrieval of doc-images within DL's; (c) the presentation of doc-images to DL users; (d) navigation within and among doc-images in DL's; and (e) effective use of personal and interactive DL's.

1 Introduction

Within digital libraries (DL's), *imaged* paper documents are growing in number and importance, but they are too often unable to play many of the useful roles that symbolically *encoded* ("born digital") documents do. Traditional document image analysis (DIA) systems can relieve some, but not all, of these obstacles. In particular, the unusually wide variety of document images found in DL's, representing many languages, historical periods, and scanning regimes, taken together pose an almost insuperable problem for present-day DIA systems. How should DIA systems be redesigned to assist in the solution of a far broader range of DIA problems than have ever been attempted before?

Section 2 summarizes the principal points relevant to this question that were aired at the International Workshop on Document Image Analysis for Libraries (DIAL2004). The issue of hardcopy books versus digital displays is raised in Section 3. Section 4 considers problems associated with document-image capture, legibility, completeness checking, support for scholarly study, and archival conservation. Certain problems arising in early-stage image processing may require fresh DIA solutions, as described in Section 5. Section 6 points out implications for DIA systems of the lack of fully automatic, high-accuracy methods for analyzing doc-image content. Needs for improved methods for presentation, display, printing, and reflowing of document images are discussed in Section 7. Retrieval, indexing, and summarization of doc-images is addressed in Section 8. Finally, Section 9 lists some problems arising in "personal" and interactive digital libraries, followed by brief conclusions in Section 10.

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 1–16, 2004.

2 The DIAL2004 Workshop

The First International Workshop on Document Image Analysis for Libraries (January 23-24, 2004, Palo Alto, CA) brought together fifty-five researchers, end-users, practitioners, business people, and end-users who were all interested in new technologies assisting the integration of imaged documents within DL's so that, ideally, everything that can be done with "born digital" data can also be done with scanned hardcopy documents. Academia, industry, and government in twelve countries were represented by researchers from the document image analysis, digital libraries, library science, information retrieval, data mining, and humanities fields. The participants worked together, in panels, debates, and group discussions, to describe the state of the art and identify urgent open problems. More broadly, the workshop attempted to stimulate closer cooperation in the future between the DIA and DL communities.

Twenty-nine regular papers, published in the proceedings [7], established the framework for discussion, which embraced six broad topics:

- DIA challenges in historical DL collections;
- handwriting recognition for DL's;
- multilingual DL's;
- DL systems architectures and costs;
- retrieval in DL's using DIA methods; and
- content extraction from document images for DL's.

The remainder of this paper summarizes work relating to these topics, with the current section placing special emphasis on the first three areas.

2.1 DIA Challenges in Historical DL Collections

Image Acquisition. Image capture from historical artifacts needs special handling to counter the defects of document aging and the physical constraints of digitization. A DIA oriented approach is suggested to effectively increase resolution and digitization speed, as well as to ensure document preservation during scanning and quality control [6, 35].

Bourgeouis *et al.* [35] use Signal to Noise Ratio (SNR) and other measures to demonstrate the loss of resolution/data in image compression formats, and recommend storage in 256 gray levels or true color. They observe that curators should be informed about the needs of DL technology and drawbacks of lossy file formats like JPEG. In addition, non-UV cold lights and automatic page turners are used to safeguard originals during scanning, and errors are countered by using skew, lighting and curvature correction for book bindings and color depth reduction for medieval documents. Character reconstruction is suggested to restore broken characters in ancient documents.

Continuous scanning is followed by automatic frame cropping as an efficient and fast procedure to generate images from microfilm [9]. Fourier-Mellin transform is used to correct rotation/shear, scale and translation errors [28]. Morphological operations, analysis of lightness and saturation in HLS (Hue, Lightness,

and Saturation) image data, and connected component analysis is used to remove reconstructed paper areas [5].

Layout Analysis and Meta-data Extraction. Layout analysis and meta-data extraction is a crucial step in creating an information base for historical DL's. Even as researchers are gaining ground on complete recognition of text content from historical documents (Subsection 2.2), practical systems have been built using only the layout analysis stage of DIA [9, 26, 35].

Availability of images makes it possible to provide content based image retrieval, using even structural features like color and layout. Marinai *et al.* [39] create an MXY tree structure during document segmentation and then use layout similarity as a feature to query documents by example.

A historical DL should supplement content with meta-data describing textual features (*e.g.*, date, author, place) and geometrical information (*e.g.*, paragraph locations, image zones). Couasnon *et al.* use an automated Web-based system for collecting annotations of French archives [18]. The system combines automatic layout analysis with human-assisted annotation in a Web interface.

Transcription of historical documents maps ASCII text to corresponding words in the document image. This is intended to circumvent the lack of perfect Optical Character Recognition (OCR) for ancient writing styles [23, 33, 66].

2.2 Handwriting Recognition for DL's

Although commercial products are available for typeset text, handwriting recognition has achieved success only in specialized domains. HMM-based character model recognizers are used in postal address recognition from mail-piece images [51, 57]. This system relies on context information related to addresses.

For transcript creation from historical documents, mapping systems use handwriting recognition. OCR engines used in these applications cannot meet real-time recognition requests. Automatic author classification systems [65] use multi-stage binarization followed by identification of document writers using character features. For Hanja scripts, OCR and UI techniques [31] incorporate nonlinear shape normalization, contour direction features and recognizers based on Mahalanobis distance to generate transcripts for Hanja (Korean) documents.

A HMM based recognizer for large lexicons is examined for indexing historical documents in [23]. The system uses substring sharing, where a prefix tree is built from the lexicon. Entries that share the same prefix also share its computation without invoking the recognizer. Duration constraints on character states, choice pruning, and parallel decoding provide a speedup of 7.7 times.

Zhang *et al.* [66] combine word model recognition and transcript mapping to create handwritten databases. Lavrenko *et al.* [34] suggest a holistic recognition technique wherein normalized word images are used as inputs to a HMM. Scalar and profile features are extracted from the images and an entire historical document is modeled as a HMM, with words constituting the state sequence. For a document written by a given author, state transition probabilities are obtained

by averaging word bi-gram probabilities collected from contemporary texts and previously transcribed writings of the target author.

2.3 Multilingual DL's

Despite excellent advances in Latin script DL's, research in other scripts such as Indic (Arabic, Bengali, Devanagari, and Telugu), Chinese, Korean, etc. is only recently receiving attention. Digital access to documents in these scripts is challenging by way of user interface (UI) design, layout analysis, and OCR.

A multilingual DL system should support simultaneous storage, entry, and display of data in many scripts. Many non-Latin scripts have a complicated character set and need a separate encoding system [17]. The display and entry of these languages requires new fonts [40, 47] and character input schemes. Also, to ensure compatibility and platform independence of data, a DL should not resort to customized solutions without completely examining existing standards.

In terms of character encoding, the Unicode Consortium aims at providing a reliable encoding scheme for all scripts in the world [17]. It currently supports all commercial scripts and is accepted as a system standard by many DL researchers and software manufacturers [11, 32, 36, 40, 60, 63]. Although alternate schemes have been suggested [43], they do not have the compatibility and global acceptance of Unicode. On the storage front, XML is emerging as a versatile and preferred scheme for DL projects [3, 32, 53, 63].

Turning to input and display techniques, multi-layered input schemes for phonetic scripts [52] are suggested for stylus/keypad based entry systems (*e.g.*, for PDA's). Keyboard mapping systems (INSCRIPT for Indic scripts) map the keys of a standard *QWERTY* keyboard onto the characters of a target script [43]. This keyboard system is functional, but has a steep learning curve. Moreover, every keyboard has to be physically labeled before a user can associate the keys with relevant characters. TrueViz [36] uses a graphical keyboard for Russian script input. Kompalli *et al.* [32] use a transliteration scheme, where Devanagari characters are entered by phonetic equivalent strings in English. For example, the Devanagari character  is entered using the English equivalent *ka*. A GUI keyboard is also provided to enter special characters.

The ability to display multiple languages on a single interface is dependent on the encoding schema and fonts used in a DL system. Most designers of multilingual software resort to Unicode-based fonts, and software vendors provide detailed guidelines for internationalization [24].

2.4 Multilingual Layout Analysis

Variation in the writing order of scripts, and the presence of language-specific constructs such as shirorekha (Devanagari), modifiers (Arabic and Devanagari), or non-regular word spacing (Arabic and Chinese) require different approaches to layout analysis. For instance, gaps may not be used to identify words in Chinese and Arabic. Techniques for script identification vary from identifying scripts of