Horst Bunke
A. Lawrence Spitz (Eds.)

LNCS 3872

Document Analysis Systems VII

7th International Workshop, DAS 2006
Nelson, New Zealand, February 2006
Proceedings

IAPR

Springer

Horst Bunke A. Lawrence Spitz (Eds.)

Document Analysis Systems VII

7th International Workshop, DAS 2006
Nelson, New Zealand, February 13-15, 2006
Proceedings

Springer

Volume Editors

Horst Bunke
University of Bern
Department of Computer Science
Neubrückstr. 10, 3012 Bern, Switzerland
E-mail: bunke@iam.unibe.ch

A. Lawrence Spitz
DocRec Ltd
34 Strathaven Place, Atawhai, Nelson 7001, New Zealand
E-mail: spitz@docrec.com

Library of Congress Control Number: 2005939178

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

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

ISSN 0302-9743
ISBN-10 3-540-32140-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-32140-8 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11669487 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3872

Lecture Notes in Computer Science

For information about Vols. 1–3771

please contact your bookseller or Springer

Vol. 3872: H. Bunke, A. L. Spitz (Eds.), Document Analysis Systems VII. XIII, 630 pages. 2006.

Vol. 3870: S. Spaccapietra, P. Atzeni, W.W. Chu, T. Catarci, K.P. Sycara (Eds.), Journal on Data Semantics V. XIII, 237 pages. 2006.

Vol. 3863: M. Kohlhase (Ed.), Mathematical Knowledge Management. XI, 405 pages. 2006. (Sublibrary LNAI).

Vol. 3861: J. Dix, S.J. Hegner (Eds.), Foundations of Information and Knowledge Systems. X, 331 pages. 2006.

Vol. 3860: D. Pointcheval (Ed.), Topics in Cryptology – CT-RSA 2006. XI, 365 pages. 2006.

Vol. 3858: A. Valdes, D. Zamboni (Eds.), Recent Advances in Intrusion Detection. X, 351 pages. 2006.

Vol. 3857: M. Fossorier, H. Imai, S. Lin, A. Poli (Eds.), Applied Algebra, Algebraic Algorithms and Error-Correcting Codes. XI, 350 pages. 2006.

Vol. 3855: E. A. Emerson, K.S. Namjoshi (Eds.), Verification, Model Checking, and Abstract Interpretation. XI, 443 pages. 2005.

Vol. 3853: A.J. Ijspeert, T. Masuzawa, S. Kusumoto (Eds.), Biologically Inspired Approaches to Advanced Information Technology. XIV, 388 pages. 2006.

Vol. 3852: P.J. Narayanan, S.K. Nayar, H.-Y. Shum (Eds.), Computer Vision - ACCV 2006, Part II. XXXI, 977 pages. 2005.

Vol. 3851: P.J. Narayanan, S.K. Nayar, H.-Y. Shum (Eds.), Computer Vision - ACCV 2006, Part I. XXXI, 973 pages. 2006.

Vol. 3850: R. Freund, G. Păun, G. Rozenberg, A. Salomaa (Eds.), Membrane Computing. IX, 371 pages. 2006.

Vol. 3848: J.-F. Boulicaut, L. De Raedt, H. Mannila (Eds.), Constraint-Based Mining and Inductive Databases. X, 401 pages. 2006. (Sublibrary LNAI).

Vol. 3847: K.P. Jantke, A. Lunzer, N. Spyratos, Y. Tanaka (Eds.), Federation over the Web. X, 215 pages. 2006. (Sublibrary LNAI).

Vol. 3844: J.-M. Bruel (Ed.), Satellite Events at the MoDELS 2005 Conference. XIII, 360 pages. 2006.

Vol. 3843: P. Healy, N.S. Nikolov (Eds.), Graph Drawing. XVII, 536 pages. 2006.

Vol. 3842: H.T. Shen, J. Li, M. Li, J. Ni, W. Wang (Eds.), Advanced Web and Network Technologies, and Applications. XXVII, 1057 pages. 2006.

Vol. 3841: X. Zhou, J. Li, H.T. Shen, M. Kitsuregawa, Y. Zhang (Eds.), Frontiers of WWW Research and Development - APWeb 2006. XXIV, 1223 pages. 2006.

Vol. 3840: M. Li, B. Boehm, L.J. Osterweil (Eds.), Unifying the Software Process Spectrum. XVI, 522 pages. 2006.

Vol. 3839: J.-C. Filliâtre, C. Paulin-Mohring, B. Werner (Eds.), Types for Proofs and Programs. VIII, 275 pages. 2006.

Vol. 3838: A. Middeldorp, V. van Oostrom, F. van Raamsdonk, R. de Vrijer (Eds.), Processes, Terms and Cycles: Steps on the Road to Infinity. XVIII, 639 pages. 2005.

Vol. 3837: K. Cho, P. Jacquet (Eds.), Technologies for Advanced Heterogeneous Networks. IX, 307 pages. 2005.

Vol. 3836: J.-M. Pierson (Ed.), Data Management in Grids. X, 143 pages. 2006.

Vol. 3835: G. Sutcliffe, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIV, 744 pages. 2005. (Sublibrary LNAI).

Vol. 3834: D.G. Feitelson, E. Frachtenberg, L. Rudolph, U. Schwiegelshohn (Eds.), Job Scheduling Strategies for Parallel Processing. VIII, 283 pages. 2005.

Vol. 3833: K.-J. Li, C. Vangenot (Eds.), Web and Wireless Geographical Information Systems. XI, 309 pages. 2005.

Vol. 3832: D. Zhang, A.K. Jain (Eds.), Advances in Biometrics. XX, 796 pages. 2005.

Vol. 3831: J. Wiedermann, G. Tel, J. Pokorný, M. Bieliková, J. Štuller (Eds.), SOFSEM 2006: Theory and Practice of Computer Science. XV, 576 pages. 2006.

Vol. 3829: P. Pettersson, W. Yi (Eds.), Formal Modeling and Analysis of Timed Systems. IX, 305 pages. 2005.

Vol. 3828: X. Deng, Y. Ye (Eds.), Internet and Network Economics. XVII, 1106 pages. 2005.

Vol. 3827: X. Deng, D.-Z. Du (Eds.), Algorithms and Computation. XX, 1190 pages. 2005.

Vol. 3826: B. Benatallah, F. Casati, P. Traverso (Eds.), Service-Oriented Computing - ICSOC 2005. XVIII, 597 pages. 2005.

Vol. 3824: L.T. Yang, M. Amamiya, Z. Liu, M. Guo, F.J. Rammig (Eds.), Embedded and Ubiquitous Computing – EUC 2005. XXIII, 1204 pages. 2005.

Vol. 3823: T. Enokido, L. Yan, B. Xiao, D. Kim, Y. Dai, L.T. Yang (Eds.), Embedded and Ubiquitous Computing – EUC 2005 Workshops. XXXII, 1317 pages. 2005.

Vol. 3822: D. Feng, D. Lin, M. Yung (Eds.), Information Security and Cryptology. XII, 420 pages. 2005.

Vol. 3821: R. Ramanujam, S. Sen (Eds.), FSTTCS 2005: Foundations of Software Technology and Theoretical Computer Science. XIV, 566 pages. 2005.

Vol. 3820: L.T. Yang, X. Zhou, W. Zhao, Z. Wu, Y. Zhu, M. Lin (Eds.), Embedded Software and Systems. XXVIII, 779 pages. 2005.

Vol. 3819: P. Van Hentenryck (Ed.), Practical Aspects of Declarative Languages. X, 231 pages. 2005.

Vol. 3818: S. Grumbach, L. Sui, V. Vianu (Eds.), Advances in Computer Science – ASIAN 2005. XIII, 294 pages. 2005.

Vol. 3817: M. Faundez-Zanuy, L. Janer, A. Esposito, A. Satue-Villar, J. Roure, V. Espinosa-Duro (Eds.), Nonlinear Analyses and Algorithms for Speech Processing. XII, 380 pages. 2006. (Sublibrary LNAI).

Vol. 3816: G. Chakraborty (Ed.), Distributed Computing and Internet Technology. XXI, 606 pages. 2005.

Vol. 3815: E.A. Fox, E.J. Neuhold, P. Premsmit, V. Wuwongse (Eds.), Digital Libraries: Implementing Strategies and Sharing Experiences. XVII, 529 pages. 2005.

Vol. 3814: M. Maybury, O. Stock, W. Wahlster (Eds.), Intelligent Technologies for Interactive Entertainment. XV, 342 pages. 2005. (Sublibrary LNAI).

Vol. 3813: R. Molva, G. Tsudik, D. Westhoff (Eds.), Security and Privacy in Ad-hoc and Sensor Networks. VIII, 219 pages. 2005.

Vol. 3810: Y.G. Desmedt, H. Wang, Y. Mu, Y. Li (Eds.), Cryptology and Network Security. XI, 349 pages. 2005.

Vol. 3809: S. Zhang, R. Jarvis (Eds.), AI 2005: Advances in Artificial Intelligence. XXVII, 1344 pages. 2005. (Sublibrary LNAI).

Vol. 3808: C. Bento, A. Cardoso, G. Dias (Eds.), Progress in Artificial Intelligence. XVIII, 704 pages. 2005. (Sublibrary LNAI).

Vol. 3807: M. Dean, Y. Guo, W. Jun, R. Kaschek, S. Krishnaswamy, Z. Pan, Q.Z. Sheng (Eds.), Web Information Systems Engineering – WISE 2005 Workshops. XV, 275 pages. 2005.

Vol. 3806: A.H. H. Ngu, M. Kitsuregawa, E.J. Neuhold, J.-Y. Chung, Q.Z. Sheng (Eds.), Web Information Systems Engineering – WISE 2005. XXI, 771 pages. 2005.

Vol. 3805: G. Subsol (Ed.), Virtual Storytelling. XII, 289 pages. 2005.

Vol. 3804: G. Bebis, R. Boyle, D. Koracin, B. Parvin (Eds.), Advances in Visual Computing. XX, 755 pages. 2005.

Vol. 3803: S. Jajodia, C. Mazumdar (Eds.), Information Systems Security. XI, 342 pages. 2005.

Vol. 3802: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part II. XLII, 1166 pages. 2005. (Sublibrary LNAI).

Vol. 3801: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part I. XLI, 1122 pages. 2005. (Sublibrary LNAI).

Vol. 3799: M. A. Rodríguez, I.F. Cruz, S. Levashkin, M.J. Egenhofer (Eds.), GeoSpatial Semantics. X, 259 pages. 2005.

Vol. 3798: A. Dearle, S. Eisenbach (Eds.), Component Deployment. X, 197 pages. 2005.

Vol. 3797: S. Maitra, C. E. V. Madhavan, R. Venkatesan (Eds.), Progress in Cryptology - INDOCRYPT 2005. XIV, 417 pages. 2005.

Vol. 3796: N.P. Smart (Ed.), Cryptography and Coding. XI, 461 pages. 2005.

Vol. 3795: H. Zhuge, G.C. Fox (Eds.), Grid and Cooperative Computing - GCC 2005. XXI, 1203 pages. 2005.

Vol. 3794: X. Jia, J. Wu, Y. He (Eds.), Mobile Ad-hoc and Sensor Networks. XX, 1136 pages. 2005.

Vol. 3793: T. Conte, N. Navarro, W.-m.W. Hwu, M. Valero, T. Ungerer (Eds.), High Performance Embedded Architectures and Compilers. XIII, 317 pages. 2005.

Vol. 3792: I. Richardson, P. Abrahamsson, R. Messnarz (Eds.), Software Process Improvement. VIII, 215 pages. 2005.

Vol. 3791: A. Adi, S. Stoutenburg, S. Tabet (Eds.), Rules and Rule Markup Languages for the Semantic Web. X, 225 pages. 2005.

Vol. 3790: G. Alonso (Ed.), Middleware 2005. XIII, 443 pages. 2005.

Vol. 3789: A. Gelbukh, Á. de Albornoz, H. Terashima-Marín (Eds.), MICAI 2005: Advances in Artificial Intelligence. XXVI, 1198 pages. 2005. (Sublibrary LNAI).

Vol. 3788: B. Roy (Ed.), Advances in Cryptology - ASIACRYPT 2005. XIV, 703 pages. 2005.

Vol. 3787: D. Kratsch (Ed.), Graph-Theoretic Concepts in Computer Science. XIV, 470 pages. 2005.

Vol. 3785: K.-K. Lau, R. Banach (Eds.), Formal Methods and Software Engineering. XIV, 496 pages. 2005.

Vol. 3784: J. Tao, T. Tan, R.W. Picard (Eds.), Affective Computing and Intelligent Interaction. XIX, 1008 pages. 2005.

Vol. 3783: S. Qing, W. Mao, J. Lopez, G. Wang (Eds.), Information and Communications Security. XIV, 492 pages. 2005.

Vol. 3782: K.-D. Althoff, A. Dengel, R. Bergmann, M. Nick, T.R. Roth-Berghofer (Eds.), Professional Knowledge Management. XXIII, 739 pages. 2005. (Sublibrary LNAI).

Vol. 3781: S.Z. Li, Z. Sun, T. Tan, S. Pankanti, G. Chollet, D. Zhang (Eds.), Advances in Biometric Person Authentication. XI, 250 pages. 2005.

Vol. 3780: K. Yi (Ed.), Programming Languages and Systems. XI, 435 pages. 2005.

Vol. 3779: H. Jin, D. Reed, W. Jiang (Eds.), Network and Parallel Computing. XV, 513 pages. 2005.

Vol. 3778: C. Atkinson, C. Bunse, H.-G. Gross, C. Peper (Eds.), Component-Based Software Development for Embedded Systems. VIII, 345 pages. 2005.

Vol. 3777: O.B. Lupanov, O.M. Kasim-Zade, A.V. Chaskin, K. Steinhöfel (Eds.), Stochastic Algorithms: Foundations and Applications. VIII, 239 pages. 2005.

Vol. 3776: S.K. Pal, S. Bandyopadhyay, S. Biswas (Eds.), Pattern Recognition and Machine Intelligence. XXIV, 808 pages. 2005.

Vol. 3775: J. Schönwälder, J. Serrat (Eds.), Ambient Networks. XIII, 281 pages. 2005.

Vol. 3774: G. Bierman, C. Koch (Eds.), Database Programming Languages. X, 295 pages. 2005.

Vol. 3773: A. Sanfeliu, M.L. Cortés (Eds.), Progress in Pattern Recognition, Image Analysis and Applications. XX, 1094 pages. 2005.

Vol. 3772: M.P. Consens, G. Navarro (Eds.), String Processing and Information Retrieval. XIV, 406 pages. 2005.

Preface

DAS 2006 is the Seventh International Association for Pattern Recognition Workshop on Document Analysis Systems and was held in Nelson, New Zealand. DAS 2006 built on the tradition of past workshops held in Kaiserslautern, Germany (1994), Malvern, PA (1996), Nagano, Japan (1998), Rio de Janeiro, Brazil (2000), Princeton, NJ (2002), and Florence, Italy (2004). The goal of this meeting is to bring together those who have designed systems, or systems components, to solve real-world problems in document analysis.

Document analysis systems is inherently an interdisciplinary field encompassing such diverse disciplines as image processing, pattern recognition, document structure and natural language processing. DAS 2006 attempted to bring these disciplines together and to provide interactions between systems developers, suppliers and end users.

We received 78 papers from 19 countries. Each submission was reviewed by three reviewers. In addition to the Program Committee members, 42 other reviewers helped in this process. From those submissions and their reviews, we went through the difficult and sometimes painful process of ranking papers for acceptance or rejection. In the end we accepted 33 papers for oral presentation and 22 for presentation at poster sessions.

We, the Co-chairmen of DAS 2006, wish to express our gratitude to all of our colleagues who have reviewed the papers submitted for this conference.

We are proud to have brought two distinguished speakers to Nelson for keynote addresses: Ian Witten of the University of Waikato, the father of the New Zealand Digital Library, and James Fruchterman, a pioneer in modern commercial optical character recognition and currently CEO of Benetech.

We owe a special debt of gratitude to Marcus Liwicki of the University of Bern for his tireless work at maintaining the website, managing the flow of papers and reviews into the ConfMan system and assembling the proceedings for publication by Springer. He was ably assisted by Andreas Schlapbach.

We are fortunate that Siemens, Hitachi and Humanware provided DAS with financial support, and we thank them for doing so. Additionally, following the DAS tradition, the organizers of DAS 2004 have passed on the surplus from running that workshop for our use.

But ultimately it is the collection of authors who submitted papers to DAS to whom we owe the greatest gratitude. It is on them and their high-quality submissions that the success of DAS 2006 relies.

February 2006

Horst Bunke and Larry Spitz
Program Chairs
DAS 2006

Organization

DAS 2006 was organized by DocRec Ltd.

Executive Committee

Conference Chairs: Larry Spitz (DocRec Ltd, New Zealand)
Horst Bunke (University of Bern, Switzerland)

Program Committee

Apostolos Antonacopoulos (UK)
Henry Baird (USA)
Thomas Breuel (Germany)
Horst Bunke (Switzerland)
Andreas Dengel (Germany)
David Doermann (USA)
Andrew Downton (UK)
Michael Fairhurst (UK)
Hiromichi Fujisawa (Japan)
Venugopal Govindaraju (USA)
Tin Kam Ho (USA)
Jianying Hu (USA)
Rolf Ingold (Switzerland)
Rangachar Kasturi (USA)
Koichi Kise (Japan)
Seong-Whan Lee (Korea)
Daniel Lopresti (USA)
Raghavan Manmatha (USA)
Simone Marinai (Italy)
Udo Miletzki (Germany)
Yasuaki Nakano (Japan)
Larry Spitz (New Zealand)
Karl Tombre (France)

Referees

Stefan Agne
Andrew Bagdanov
Ardhendu Behera
Koustav Bhattacharya
Alain Biem
Jean-Luc Blöchle
Matthew Boonstra
Jakob Brendel
Joshua Candamo
Farzin Deravi

Faisal Farooq
Gunnar Grimnes
Richard Guest
Sanaul Hoque
Gareth Howells
Jonathan Hull
Masakazu Iwamura
Stefan Jaeger
Thomas Kieninger
Malte Kiesel
Bertin Klein
Dar-Shyang Lee
Hansheng Lei
Jian Liang
Rainer Lindwurm
Vasant Manohar
Dalila Mekhaldi
David Mihalcik
Tristan Miller
Pranab Mohanty
Sunita Nayak
Shinichiro Omachi
Christoph Pesch
Maurizio Rigamonti
Thomas Roth-Berghofer
Sven Schwarz
Karthik Sridharan
Seiichi Uchida
Himanshu Vajaria
Ludger van Elst
Shankar Vembu
Alan Yang

Sponsoring Institutions

Siemens AG, Munich, Germany
HumanWare Group, Christchurch, New Zealand
Hitachi Central Research Laboratory, Tokyo, Japan

Scientific Sponsors

DocRec Ltd, Atawhai, Nelson, New Zealand
University of Bern, Switzerland
International Association for Pattern Recognition

Table of Contents

Session 1: Digital Libraries

Session 2: Image Processing

Session 3: Handwriting 1

Session 4: Document Structure and Format

Session 5: Tables

Session 6: Handwriting 2

Session 7: Language and Script Identification

Session 9: Systems and Performance Evaluation

Session 10: Retrieval and Segmentation

Posters

Retrieval from Document Image Collections

A. Balasubramanian, Million Meshesha, and C.V. Jawahar

Centre for Visual Information Technology,
International Institute of Information Technology,
Hyderabad - 500 032, India
jawahar@iiit.ac.in

Abstract. This paper presents a system for retrieval of relevant documents from large document image collections. We achieve effective search and retrieval from a large collection of printed document images by matching image features at word-level. For representations of the words, profile-based and shape-based features are employed. A novel DTW-based partial matching scheme is employed to take care of morphologically variant words. This is useful for grouping together similar words during the indexing process. The system supports cross-lingual search using OM-Trans transliteration and a dictionary-based approach. System-level issues for retrieval (eg. scalability, effective delivery etc.) are addressed in this paper.

1 Introduction

Large digital libraries, such as Digital Library of India (DLI) [1] are emerging for archiving large collection of printed and handwritten documents. The DLI aims at digitizing all literary, artistic, and scientific works of mankind so as to create better access to traditional materials, easier preservation, and make documents freely accessible to the global society. More than 25 scanning centers all over India are working on digitization of books and manuscripts. The mega scanning center we have, has around fifty scanners, each one of them capable lof scanning approximately 5000 pages in 8 hours. As on September 2005, close to 100 thousand books with 25 million pages were digitized and made available online by DLI (*http://dli.iiit.ac.in*) as document images.

Building an effective access to these document images requires designing a mechanism for effective search and retrieval of textual data from document image collections. Document image indexing and retrieval were studied with limited scope in literature [2]. Success of these procedures mainly depends on the performance of the OCRs, which convert the document images into text. Much of the data in DLI are in Indian languages. Searching in these document image collections based on content, is not presently possible. This is because OCRs are not yet able to successfully recognize printed texts in Indian languages. We need an alternate approach to access the content of these documents [3]. A promising alternate direction is to search for relevant documents in image space without any explicit recognition. We have been motivated by the successful attempts on

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 1–12, 2006.

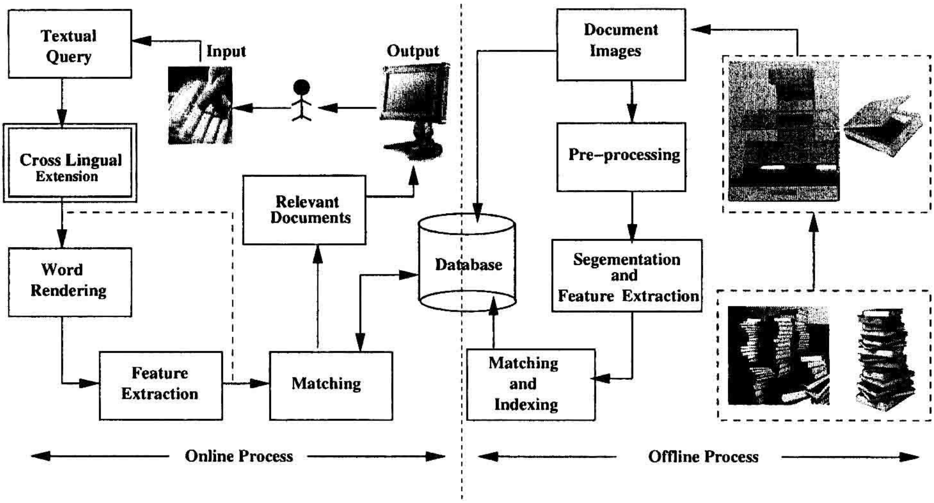

Fig. 1. Conceptual Diagram of the Searching Procedure from Multilingual Document Image Database. A Web Demo for the Above Procedure is Available Online at http://cvit.iiit.ac.in/wordsearch.html.

locating a specific word in handwritten English documents by matching image features for historical documents [4, 5].

We have already addressed algorithmic challenges for effective search in document images [6] . This paper describes the issues associated with the implementation of a scalable system for Indian language document images. A conceptual block diagram of our prototype system is shown in Figure 1. Our system accepts textual query from users. The textual query is first converted to an image by rendering, features are extracted from these images and then search is carried out for retrieval of relevant documents. Results of the search are pages from document image collections containing queried word sorted based on their relevance to the query.

2 Challenges in Design and Implementation of the System

Search and retrieval from large collection of document images is a challenging task, specially when there is no textual representation available. To design and implement a successful search engine in image domain, we need to address the following issues.

Search in images: Search in image space requires appropriate representational schemes and similarity measures. Success of content-based image retrieval(CBIR) schemes were limited by the diversity of the image collections. Digital libraries primarily archive text images, but of varying quality, script, style, size and font.

We need to come up with appropriate features and matching schemes, which can represent the content (text), while invariant to the popular variations.

Degradations of documents: Documents in digital libraries are extremely poor in quality. Popular artifacts in printed document images include (a) Excessive dusty noise, (b) Large ink-blobs joining disjoint characters or components, (c) Vertical cuts due to folding of the paper, (d) Cuts of arbitrary direction due to paper quality or foreign material, (e) Degradation of printed text due to the poor quality of paper and ink, (f) Floating ink from facing pages etc. We need to design an appropriate representational scheme and matching algorithm to accommodate the effect of degradation.

Need for cross-lingual retrieval: Document images in digital libraries are from diverse languages. Relevant documents that users need may be available in different languages. Most educated Indians can read more than one language. Hence, we need to design a mechanism that allows users to retrieve all documents related to their queries in any of the Indian languages.

Computational speed: Searching from large collection of document images pass through many steps: image processing, feature extraction, matching and retrieval of relevant documents. Each of these steps could be computationally expensive. In a typical book, there could be around 90,000 words and processing all of them online is practically impossible. We do all computationally expensive operations during the offline indexing (Section 4) and do minimal operations during online retrieval (Section 5).

Indian languages: Indian languages pose many additional challenges [7]. Some of these are: (i) lack of standard representation for the fonts and encoding, (ii) lack of support from operating system, browsers and keyboard, and (iii) lack of language processing routines. These issues add to the complexity of the design and implementation of a document image retrieval system.

3 Representation and Matching of Word Images

Word images extracted from documents in digital libraries are of varying quality, script, font, size and style. An effective representation of the word images will have to take care of these artifacts for successful searching and retrieval. We combined two categories of features to address these effects: word profiles and structural features. Explicit definitions of these features may be seen in [6].

Word Profiles: Profiles of the word provide a coarse way of representing a word image for matching. Profiles like upper word, lower word, projection and transition profiles are used here for word representation. Upper and lower word profiles capture part of the outlining shape of a word, while projection and transition profiles capture the distribution of ink along one of the two dimensions in a word image.