Proceedings

Third Annual Symposium on Document Analysis and Information Retrieval

April 11 - 13, 1994

Las Vegas, Nevada

Sponsored by:

Information Science Research Institute

and

Howard R. Hughes College of Engineering

University of Nevada, Las Vegas
4505 Maryland Parkway • Box 454021
Las Vegas, Nevada 89154-4021
# Table of Contents

Conference Chairman's Message ................................................................. i  
Message From The President, University of Nevada, Las Vegas ................ ii  
Message From The Dean, College of Engineering ........................................ iii  
Board of Advisors, Information Science Research Institute ...................... iv  
Conference Committee Members ................................................................... v

## INVITED PAPERS

Recognition Error and Reject Trade-off ....................................................... 1  
   C. K. Chow

Text Retrieval Using the Vector Processing Model ....................................... 9  
   Gerard Salton and James Allan

Computational Geometry for Document Analysis ......................................... 23  
   Godfried T. Toussaint

The Text Retrieval Conference (TREC) ....................................................... 43  
   Donna Harman

## CONTRIBUTED PAPERS

Adaptive Logic Networks for Machine-Printed Character Recognition .......... 69  

A Comparison of Two Learning Algorithms for Text Categorization .......... 81  
   David D. Lewis and Marc Ringuette

Optimizing Parameters in a Ranked Retrieval System Using Multi-Query  
   Relevance Feedback ............................................................................... 95  
   Brian T. Bartell, Garrison W. Cottrell and Richard K. Belew

Pattern Classification Based on Adaptive Self-Organizing Neural Network . . 103  
   Yuan Y. Tang, Ching Y. Suen, Tao Li and L. Y. Fang

An Evaluation of Information Retrieval Accuracy with Simulated OCR Output .115  
   W. B. Croft, S. M. Harding, K. Taghva and J. Borsack

Validation of Simulated OCR Data Sets ..................................................... 127  
   George Nagy

Validation of Document Image Defect Models for Optical Character Recognition .137  
   Yanhong Li, Daniel Lopresti and Andrew Tomkins
Keyword Selection from Word Recognition Results using Definitional Overlap. 151
Paul Filipski and Jonathan J. Hull

N-Gram-Based Text Categorization ................................................. 161
William B. Cavnar and John M. Trenkle

Lexicon-Based Word Recognition Without Word Segmentation ................. 177
Gregory K. Myers and Chien-Huei Chen

Automatic Indexing of Compound Words based on Mutual Information for
Korean Text Retrieval ................................................................. 189
Pan Koo Kim and Yoo Kun Cho

The Role of Visualization in Document Analysis ................................ 199
Robert R. Korfhage and Kai A. Olsen

About the Logical Partitioning of Document Images ............................. 209
Andreas Dengel

A Context-Based Approach to Text Recognition ................................ 219
T. G. Rose, L. J. Evett and A. C. Jobbins

Script and Language Determination from Document Images .................. 229
A. Lawrence Spitz

Binarization and Multi-Thresholding of Document Images Using Connectivity 237
Lawrence O'Gorman

Direct Extraction of Topographic Features from Gray Scale Character Images ... 253
Seong-Whan Lee and Young Joon Kim

An Alternative to Vectorization: Decomposition of Graphics Into Primitives ... 263
J. E. den Hartog, T. K. ten Kate, J. J. Gerbrands and G. van Antwerpen

Asymptotic Accuracy of Two-Class Discrimination ................................ 275
Tin Kam Ho and Henry S. Baird

Use of Constraints as a Second Stage Character Classification Technique ...... 289
George Sazaklis and Theo Pavlidis

Performance Evaluation of Two OCR Systems .................................... 299
Su Chen, Suresh Subramaniam, Robert M. Haralick and Ihsin T. Phillips

An Approach to Interactive Retrieval in Face Image Databases Based on
Semantic Attributes ........................................................................ 319
Venkat N. Gudivada, Vijay V. Raghavan and Guna S. Seetharaman
Marking of Document Images with Codewords to Deter Illicit Dissemination
J. T. Brassil, S. Low, N. F. Maxemchuk and L. O'Gorman

Modelling and Exploiting Traceability Between Development Documents
J. P. Queille, A. Richermo, J. F. Voidrot and F. Sedes

Development of a Full-Text Information Retrieval System
Keizo Oyama, Akira Miyazawa, Atsuhiro Takasu and Kouji Shibano

POSTER PAPERS

Low-level Structural Recognition of Documents
Y. Chenevoy and A. Belaid

Document Characterization, Authentication and Retrieval Based on
Medium-Embedded Random Patterns
Dragana Brzakovic and Nenad Vujovic

Information Retrieval for a Document Writing Assistance System
Marie-Louise Corral, Christine Julien and Amaury Simon

Issues in Automatic OCR Error Classification
Jeffrey Esakov, Daniel P. Lopresti, Jonathan S. Sandberg and Jiangying Zhou

Correlated Run Length Algorithm (CURL) for Detecting Form Structure
within Digitized Documents
Michael D. Garris

Experiments in Automatic Word Class and Word Sense Identification for
Information Retrieval
Susan Gauch and Robert P. Futrelle

Estimating Errors in Document Databases
Jaekyu Ha, Robert M. Haralick, Su Chen and Ihsin T. Phillips

Retrieval of Line Drawings
Oliver Lorenz and Gladys Monagan

A Paper Form Processing System with an Error Correcting Function for
Reading Handwritten Kanji Strings
Katsumi Marukawa, Kazuki Nakashima, Masashi Koga, Yoshihiro Shima
and Hiromichi Fujisawa

Author Index