Synergies from Document Analysis

Daniel P. Lopresti

lopresti@cse.lehigh.edu

Professor, Department of Computer Science and Engineering, Lehigh University

President, International Association for Pattern Recognition (IAPR)

Vice Chair, Computing Research Association's Computing Community Consortium (CCC)

Goals

- Provide short overview of document analysis research.*
- Survey history of the field briefly.
- Describe some current problems of interest.
- Highlight potential synergies with forensic science.
- Point to online resources that may be useful.
- Offer my thanks to Samiah Ibrahim, ASQDE President, and other conference organizers for providing this opportunity.

* Cannot completely avoid discussing handwriting, but will try to minimize potential duplication.

My background

- Professor of Computer Science & Engineering at Lehigh.
- President of International Association for Pattern Recognition.
- Vice Chair of Computing Research Association’s CCC Council.
- 30 years in the field; co-EIC of IJDAR; co-Program Chair of ICDAR 2021 and numerous other past conferences.
- B.A. from Dartmouth, Ph.D. from Princeton.
- Research interests include algorithmic and systems-related questions in document analysis, pattern recognition / machine learning, and computer security including electronic voting.

What is document analysis?

Documents are one of humankind’s most significant creations:

- Attempt to provide machines with human levels of capability when it comes interpreting documents, broadly defined.

Document analysis combines:

- Computer vision.
- Image processing.
- Machine learning (pattern recognition).
- Natural language understanding.
- Domain expertise.

What is the target of document analysis?

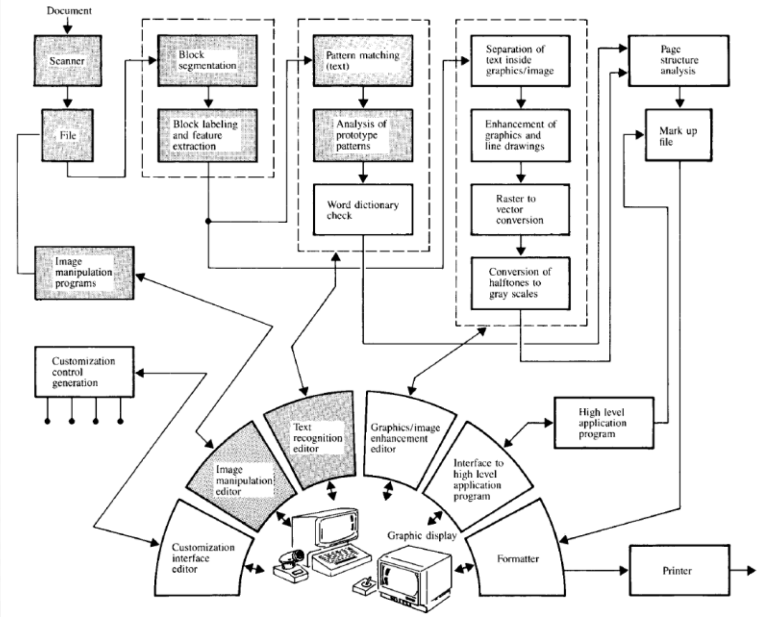
plain text	correct word order for OCR
illustrated text	reading order, links to illustrations
structured text	compilable or executable form
envelope, letter	routing information
directory, TOC	name-attribute pairs
business form	links to database, add tags
schematic diagram	net list or graph
engineering drawing	current CAD format
map	GIS representation
music score	MIDI representation
table	layout-independent descriptor

Adapted from “State of Art of Document Image Processing” (PowerPoint presentation), G. Nagy.

History of document analysis: a few highlights

- OCR – optical character recognition – dates back to early 1900's.
- Growing practical interest in 1970's (Ray Kurzweil and others).

From a
1982 paper:

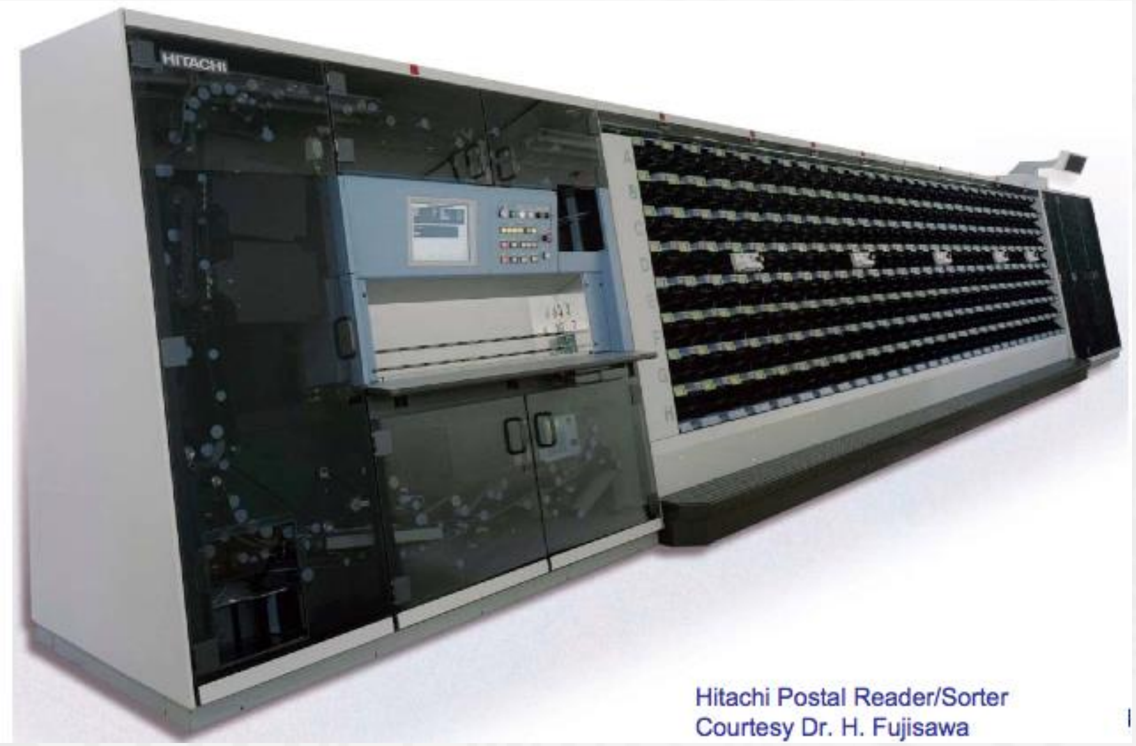

Document
Scanner
File
Block segmentation
Block labeling and feature extraction
Pattern matching (text)
Analysis of prototype patterns
Word dictionary check
Separation of text inside graphics/image
Enhancement of graphics and line drawings
Raster to vector conversion
Conversion of halftones to gray scales
Page structure analysis
Mark up file
Image manipulation programs
Customization control generation
High level application program
Customization interface editor
Image manipulation editor
Text recognition editor
Graphics/image enhancement editor
Interface to high level application program
Graphic display
Formatter
Printer

- Block segmentation
- OCR
- Dictionary check
- Graphics ID
- Vectorization
- Halftone to grayscale
- Page markup
- Various editors

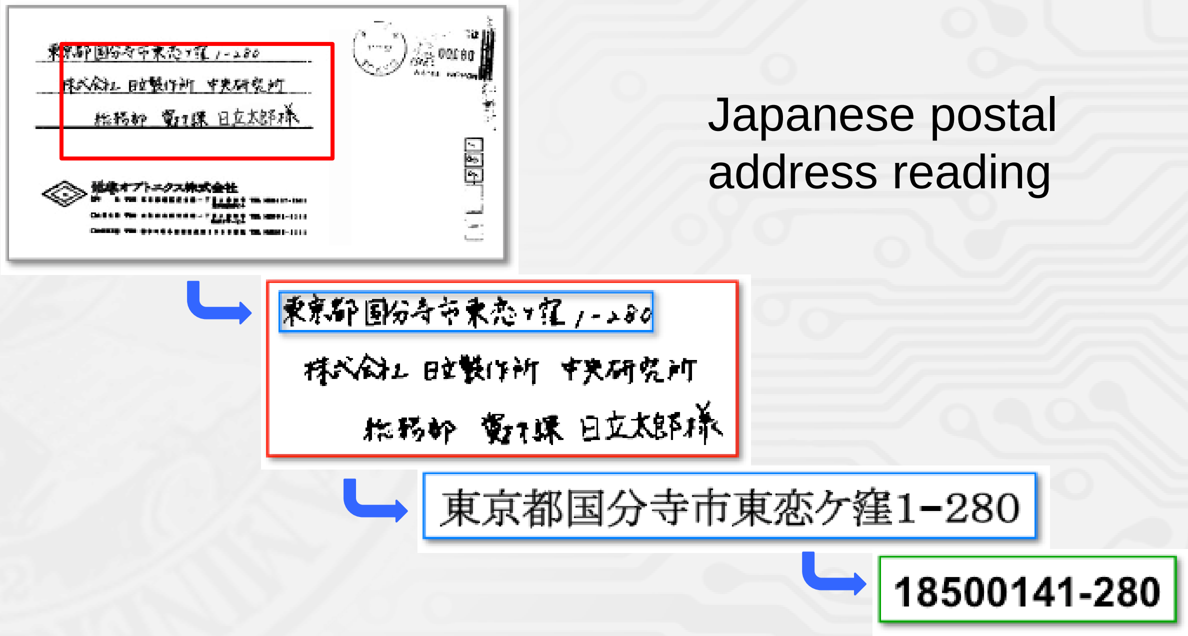
"Document Analysis System," K. Y. Wong, R. G. Casey, and F. M. Wahl, IBM Journal of Research and Development, vol. 26, no. 6, November 1982, pp. 647-656.

History of document analysis: a few highlights

Hitachi high-speed postal address reader from 2006:



Hitachi Postal Reader/Sorter
Courtesy Dr. H. Fujisawa



東京都国分寺市東恋ケ窪1−280

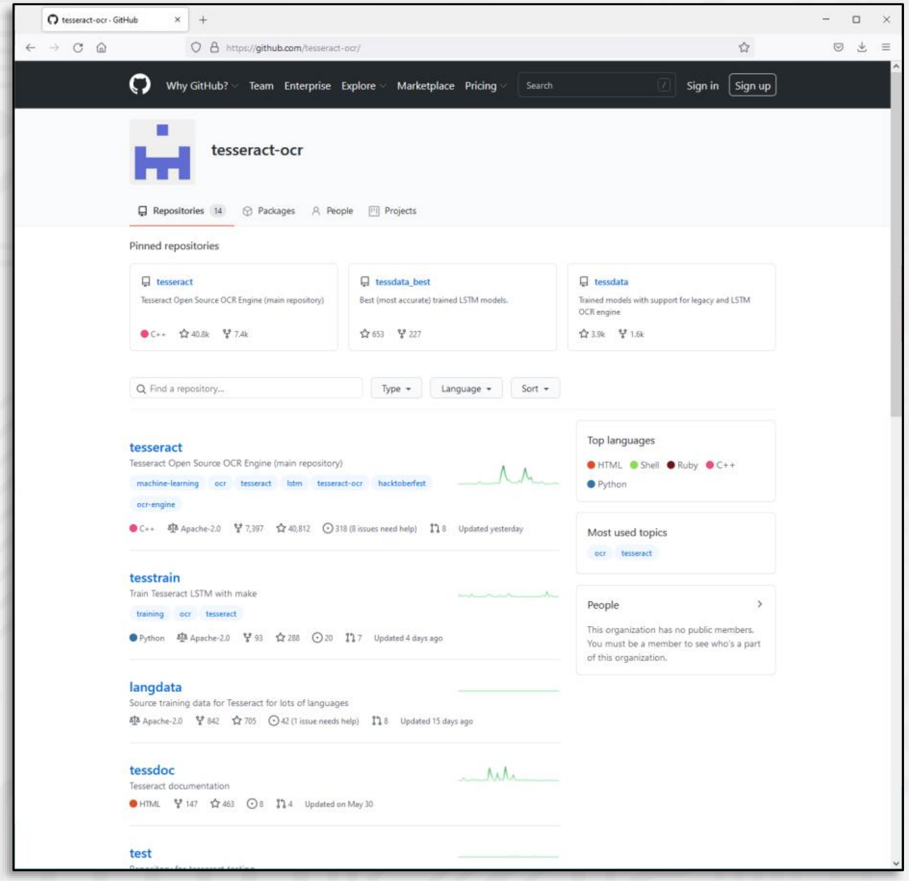
18500141-280

Japanese postal address reading

“How to Deal with Uncertainty and Variability: Experience and Solutions” (PowerPoint presentation), H. Fujisawa, 2006.

History of document analysis: a few highlights

Open Source:


tesseract-ocr · GitHub
https://github.com/tesseract-ocr/
Why GitHub? Team Enterprise Explore Marketplace Pricing
Search
Sign in
Sign up

tesseract-ocr

Repositories 14 · Packages · People · Projects

Pinned repositories

tesseract Tesseract Open Source OCR Engine (main repository) ● C++ ☆ 40.8k 7.4k	tessdata_best Best (most accurate) trained LSTM models. ☆ 653 227	tessdata Trained models with support for legacy and LSTM OCR engine ☆ 3.9k 1.6k

Find a repository... · Type ▾ · Language ▾ · Sort ▾

tesseract
Tesseract Open Source OCR Engine (main repository)
machine-learning · ocr · tesseract · lstm · tesseract-ocr · hacktoberfest · ocr-engine
● C++ · Apache-2.0 · 7,397 · ☆ 40,812 · 318 (8 issues need help) · 8 · Updated yesterday

testrain
Train Tesseract LSTM with make
training · ocr · tesseract
● Python · Apache-2.0 · 93 · ☆ 288 · 20 · 7 · Updated 4 days ago

langdata
Source training data for Tesseract for lots of languages
Apache-2.0 · 842 · ☆ 705 · 42 (1 issue needs help) · 8 · Updated 15 days ago

tessdoc
Tesseract documentation
● HTML · 147 · ☆ 463 · 8 · 4 · Updated on May 30

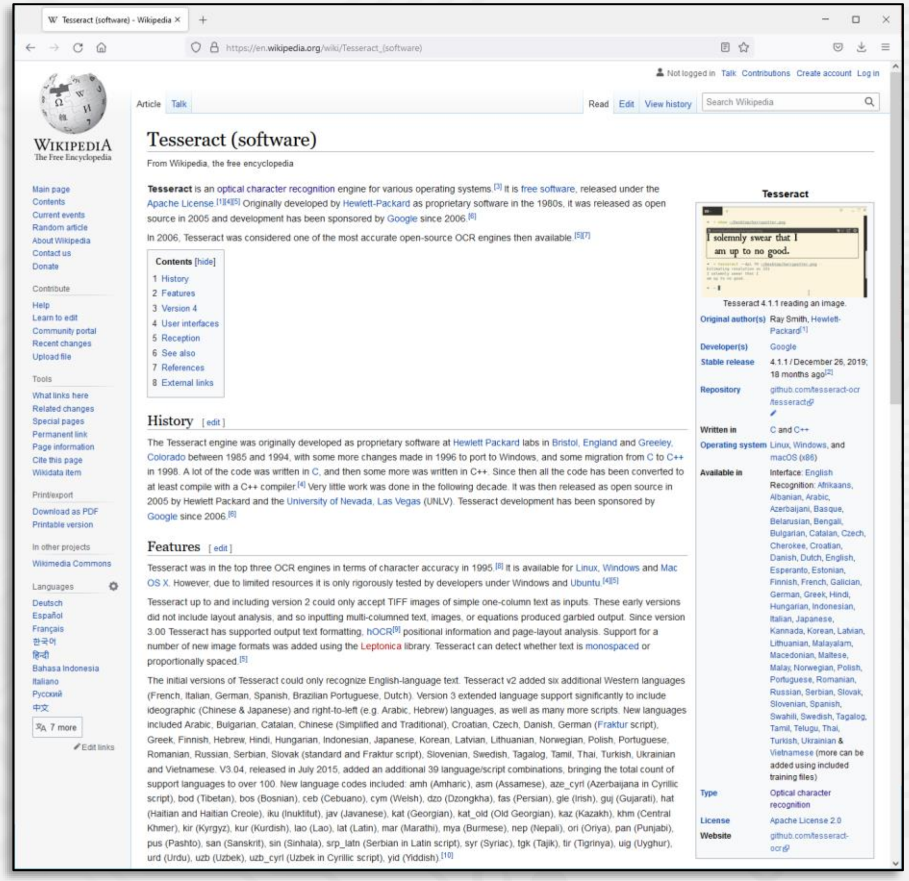
test

Top languages
● HTML ● Shell ● Ruby ● C++ ● Python

Most used topics
ocr · tesseract

People
This organization has no public members. You must be a member to see who's a part of this organization.

Tesseract OCR



Tesseract (software)

From Wikipedia, the free encyclopedia

Tesseract is an optical character recognition engine for various operating systems.[3] It is free software, released under the Apache License.[1][4][5] Originally developed by Hewlett-Packard as proprietary software in the 1980s, it was released as open source in 2005 and development has been sponsored by Google since 2006.[6]

In 2006, Tesseract was considered one of the most accurate open-source OCR engines then available.[5][7]

Contents [hide]

1 History
2 Features
3 Version 4
4 User interfaces
5 Reception
6 See also
7 References
8 External links

History [edit]

The Tesseract engine was originally developed as proprietary software at Hewlett Packard labs in Bristol, England and Greeley, Colorado between 1985 and 1994, with some more changes made in 1996 to port to Windows, and some migration from C to C++ in 1998. A lot of the code was written in C, and then some more was written in C++. Since then all the code has been converted to at least compile with a C++ compiler.[4] Very little work was done in the following decade. It was then released as open source in 2005 by Hewlett Packard and the University of Nevada, Las Vegas (UNLV). Tesseract development has been sponsored by Google since 2006.[6]

Features [edit]

Tesseract was in the top three OCR engines in terms of character accuracy in 1995.[8] It is available for Linux, Windows and Mac OS X. However, due to limited resources it is only rigorously tested by developers under Windows and Ubuntu.[4][5]

Tesseract up to and including version 2 could only accept TIFF images of simple one-column text as inputs. These early versions did not include layout analysis, and so inputting multi-columned text, images, or equations produced garbled output. Since version 3.00 Tesseract has supported output text formatting, hOCR[9] positional information and page-layout analysis. Support for a number of new image formats was added using the Leptonica library. Tesseract can detect whether text is monospaced or proportionally spaced.[5]

The initial versions of Tesseract could only recognize English-language text. Tesseract v2 added six additional Western languages (French, Italian, German, Spanish, Brazilian Portuguese, Dutch). Version 3 extended language support significantly to include ideographic (Chinese & Japanese) and right-to-left (e.g. Arabic, Hebrew) languages, as well as many more scripts. New languages included Arabic, Bulgarian, Catalan, Chinese (Simplified and Traditional), Croatian, Czech, Danish, German (Fraktur script), Greek, Finnish, Hebrew, Hindi, Hungarian, Indonesian, Japanese, Korean, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak (standard and Fraktur script), Slovenian, Swedish, Tagalog, Tamil, Thai, Turkish, Ukrainian and Vietnamese. V3.04, released in July 2015, added an additional 39 language/script combinations, bringing the total count of support languages to over 100. New language codes included: amh (Amharic), asm (Assamese), aze_cyrl (Azerbaijana in Cyrillic script), bod (Tibetan), bos (Bosnian), ceb (Cebuano), cym (Welsh), dzo (Dzongkha), fas (Persian), gle (Irish), guj (Gujarati), hat (Haitian and Haitian Creole), iku (Inuktitut), jav (Javanese), kat (Georgian), kat_old (Old Georgian), kaz (Kazakh), khm (Central Khmer), kir (Kyrgyz), kur (Kurdish), lao (Lao), lat (Latin), mar (Marathi), mya (Burmese), nep (Nepali), ori (Oriya), pan (Punjabi), pus (Pashto), san (Sanskrit), sin (Sinhala), srp_latn (Serbian in Latin script), syr (Syriac), tgk (Tajik), tir (Tigrinya), uig (Uyghur), urd (Urdu), uzb (Uzbek), uzb_cyrl (Uzbek in Cyrillic script), yid (Yiddish).[10]

Tesseract

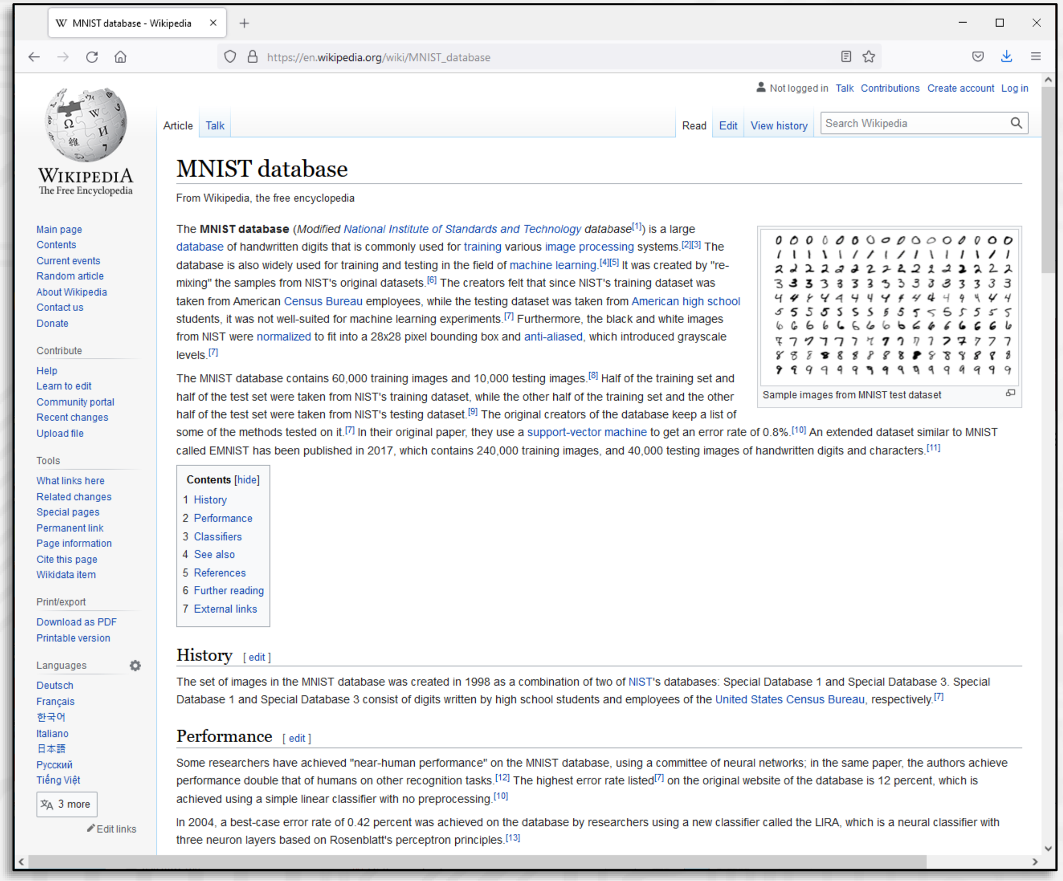
I solemnly swear that I am up to no good.

Tesseract 4.1.1 reading an image.

Original author(s)	Ray Smith, Hewlett-Packard[1]
Developer(s)	Google
Stable release	4.1.1 / December 26, 2019; 18 months ago[2]
Repository	github.com/tesseract-ocr/tesseract
Written in	C and C++
Operating system	Linux, Windows, and macOS (x86)
Available in	Interface: English Recognition: Afrikaans, Albanian, Arabic, Azerbaijani, Basque, Belarusian, Bengali, Bulgarian, Catalan, Czech, Cherokee, Croatian, Danish, Dutch, English, Esperanto, Estonian, Finnish, French, Galician, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Kannada, Korean, Latvian, Lithuanian, Malayalam, Macedonian, Maltese, Malay, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swahili, Swedish, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian & Vietnamese (more can be added using included training files)
Type	Optical character recognition
License	Apache License 2.0
Website	github.com/tesseract-ocr

History of document analysis: a few highlights

Seeded Deep Learning revolution:



MNIST database

From Wikipedia, the free encyclopedia

The **MNIST database** (*Modified National Institute of Standards and Technology database*[1]) is a large database of handwritten digits that is commonly used for training various image processing systems.[2][3] The database is also widely used for training and testing in the field of machine learning.[4][5] It was created by "re-mixing" the samples from NIST's original datasets.[6] The creators felt that since NIST's training dataset was taken from American Census Bureau employees, while the testing dataset was taken from American high school students, it was not well-suited for machine learning experiments.[7] Furthermore, the black and white images from NIST were normalized to fit into a 28x28 pixel bounding box and anti-aliased, which introduced grayscale levels.[7]

Sample images from MNIST test dataset

The MNIST database contains 60,000 training images and 10,000 testing images.[8] Half of the training set and half of the test set were taken from NIST's training dataset, while the other half of the training set and the other half of the test set were taken from NIST's testing dataset.[9] The original creators of the database keep a list of some of the methods tested on it.[7] In their original paper, they use a support-vector machine to get an error rate of 0.8%.[10] An extended dataset similar to MNIST called EMNIST has been published in 2017, which contains 240,000 training images, and 40,000 testing images of handwritten digits and characters.[11]

Contents [hide]

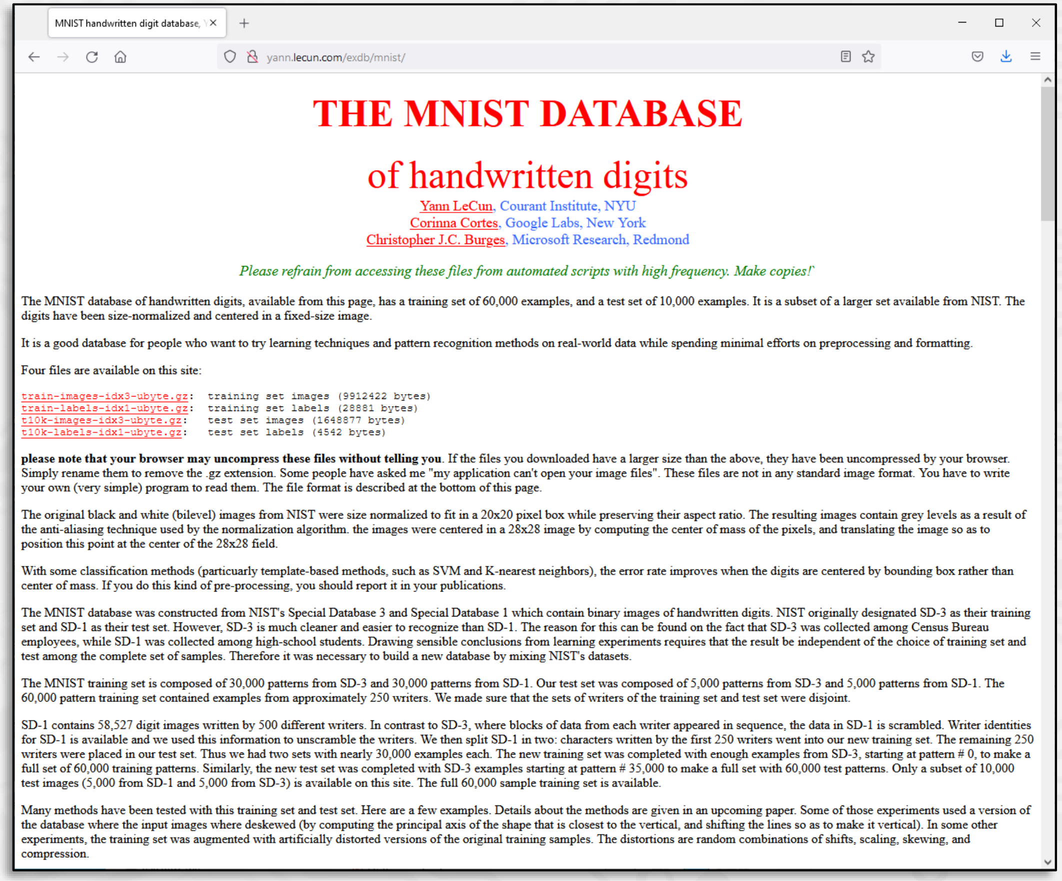
History [edit]

The set of images in the MNIST database was created in 1998 as a combination of two of NIST's databases: Special Database 1 and Special Database 3. Special Database 1 and Special Database 3 consist of digits written by high school students and employees of the United States Census Bureau, respectively.[7]

Performance [edit]

Some researchers have achieved "near-human performance" on the MNIST database, using a committee of neural networks; in the same paper, the authors achieve performance double that of humans on other recognition tasks.[12] The highest error rate listed[7] on the original website of the database is 12 percent, which is achieved using a simple linear classifier with no preprocessing.[10]

In 2004, a best-case error rate of 0.42 percent was achieved on the database by researchers using a new classifier called the LIRA, which is a neural classifier with three neuron layers based on Rosenblatt's perceptron principles.[13]



THE MNIST DATABASE

of handwritten digits

Yann LeCun, Courant Institute, NYU
Corinna Cortes, Google Labs, New York
Christopher J.C. Burges, Microsoft Research, Redmond

Please refrain from accessing these files from automated scripts with high frequency. Make copies!

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Four files are available on this site:

```
train-images-idx3-ubyte.gz:  training set images (9912422 bytes)
train-labels-idx1-ubyte.gz:  training set labels (28881 bytes)
t10k-images-idx3-ubyte.gz:   test set images (1648877 bytes)
t10k-labels-idx1-ubyte.gz:   test set labels (4542 bytes)
```

please note that your browser may uncompress these files without telling you. If the files you downloaded have a larger size than the above, they have been uncompressed by your browser. Simply rename them to remove the .gz extension. Some people have asked me "my application can't open your image files". These files are not in any standard image format. You have to write your own (very simple) program to read them. The file format is described at the bottom of this page.

The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. the images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field.

With some classification methods (particuarly template-based methods, such as SVM and K-nearest neighbors), the error rate improves when the digits are centered by bounding box rather than center of mass. If you do this kind of pre-processing, you should report it in your publications.

The MNIST database was constructed from NIST's Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and SD-1 as their test set. However, SD-3 is much cleaner and easier to recognize than SD-1. The reason for this can be found on the fact that SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-school students. Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples. Therefore it was necessary to build a new database by mixing NIST's datasets.

The MNIST training set is composed of 30,000 patterns from SD-3 and 30,000 patterns from SD-1. Our test set was composed of 5,000 patterns from SD-3 and 5,000 patterns from SD-1. The 60,000 pattern training set contained examples from approximately 250 writers. We made sure that the sets of writers of the training set and test set were disjoint.

SD-1 contains 58,527 digit images written by 500 different writers. In contrast to SD-3, where blocks of data from each writer appeared in sequence, the data in SD-1 is scrambled. Writer identities for SD-1 is available and we used this information to unscramble the writers. We then split SD-1 in two: characters written by the first 250 writers went into our new training set. The remaining 250 writers were placed in our test set. Thus we had two sets with nearly 30,000 examples each. The new training set was completed with enough examples from SD-3, starting at pattern # 0, to make a full set of 60,000 training patterns. Similarly, the new test set was completed with SD-3 examples starting at pattern # 35,000 to make a full set with 60,000 test patterns. Only a subset of 10,000 test images (5,000 from SD-1 and 5,000 from SD-3) is available on this site. The full 60,000 sample training set is available.

Many methods have been tested with this training set and test set. Here are a few examples. Details about the methods are given in an upcoming paper. Some of those experiments used a version of the database where the input images where deskewed (by computing the principal axis of the shape that is closest to the vertical, and shifting the lines so as to make it vertical). In some other experiments, the training set was augmented with artificially distorted versions of the original training samples. The distortions are random combinations of shifts, scaling, skewing, and compression.

More recent topics (from ICDAR 2021 CFP)

- Document image processing
- Text and symbol recognition
- Document analysis systems
- Indexing and retrieval of documents
- Extracting document semantics
- Document summarization and translation
- Human document interaction
- Mobile text recognition
- Scene text detection and recognition
- Recognition of tables and formulas
- Signature verification
- Medical document analysis
- Document analysis for literature search
- Physical and logical layout analysis
- Handwriting recognition
- Document classification
- Document synthesis
- NLP for document understanding
- Office automation
- Multimedia document analysis
- Pen-based document analysis
- Graphics recognition
- Historical document analysis
- Document forensics and provenance analysis
- Document analysis for social good
- Gold-standard benchmarks and data sets

Is document analysis primarily retrospective?

- Most documents we wish to keep are now produced digitally: books, journals, newspapers, letters, drawings, forms (like tax returns and visa applications).

But …

- Many pre-1980 documents remain to be converted, some of business value (utility drawings), and many historical artifacts.
- Original software or digital medium is not always available (conversion of CAD drawings, tech journals, census data).
- Digital version is not always available ⇒ personal applications.

Adapted from “State of Art of Document Image Processing” (PowerPoint presentation), G. Nagy.

Popular venues: ICDAR

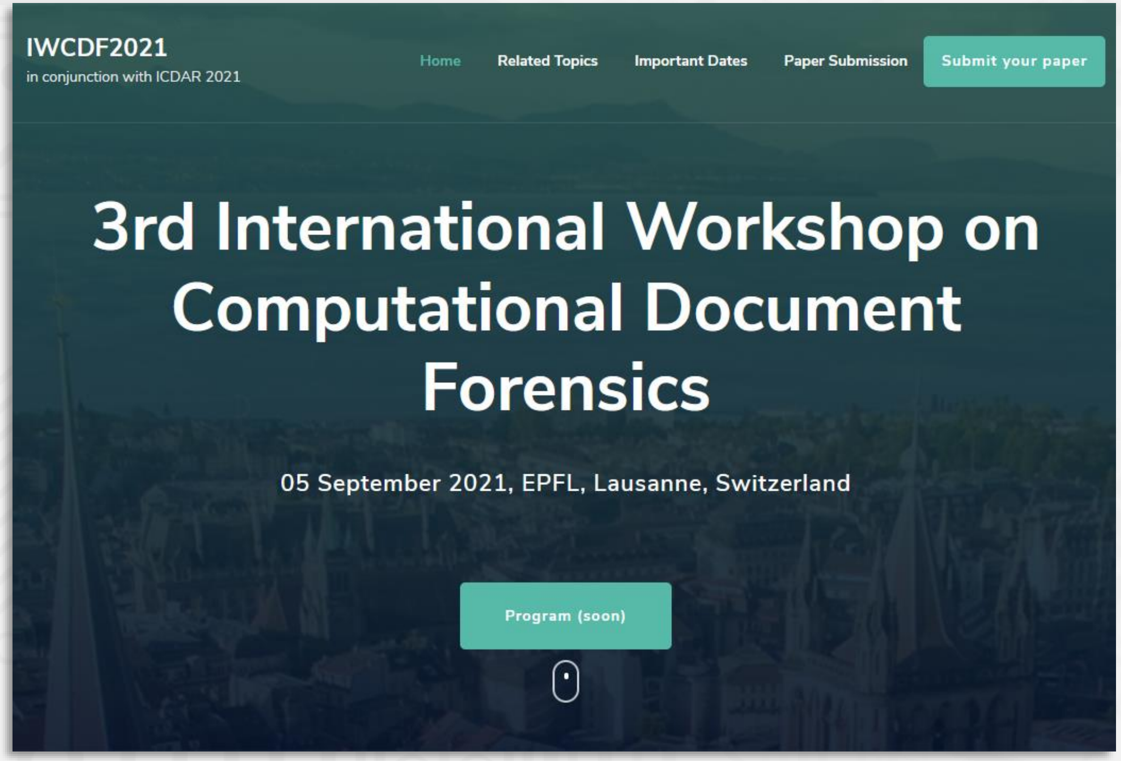
- The International Conference on Document Analysis and Recognition has been held every two years since 1991. Next conference will be held in Lausanne, Switzerland in September.

16th International Conference on Document Analysis and Recognition
ICDAR 2021
September 5-10, 2021, Lausanne, Switzerland

- Biggest, broadest conference in document analysis.
- ~200 papers, ~400 attendees.
- Also workshops, competitions, tutorials.

Popular venues: IWCDF

- ICDAR 2021 Workshop on Computational Document Forensics.



IWCDF2021

in conjunction with ICDAR 2021

3rd International Workshop on Computational Document Forensics

05 September 2021, EPFL, Lausanne, Switzerland

“Everywhere around the world, industries and government processes are being more and more digitized. Document management systems and digital safe-boxes are particularly concerned by these questions, since documents generally remain the basis of many decisions for transactions, contracts and communication. Documents also remain the proofs for many legal issues. As a consequence, it becomes absolutely essential to develop computational forensic science applied to documents and to create the conditions for protecting documents, for confirming their authenticity and for detecting frauds.”

Popular venues: IWCDF

- Prevention of forgeries in documents
- Detection of forged documents
- Detection of fake documents
- Authentication of documents
- Forgery localisation
- Copyright protection
- Watermarking
- Digital signatures
- Taxonomy of features
- Expert results vs. system outputs

- Forensic handwriting verification/identification
- Forensic signature verification/identification
- Within writer versus between writer variations
- Determining the frequency of occurrence of handwriting features
- Automated signature identification and verification
- Automated handwriting identification and verification
- Extraction of movement order features out of the ink trace
- Allograph matching and clustering
- Classification of signatures: legible vs. illegible, complex vs. simple
- Detection of forgeries in printed and rescanned documents

Popular venue: DAS

- The Workshop on Document Analysis Systems (DAS) has been held every two years since 1996, most recently virtually in July 2020 (was to have been in Wuhan, China).

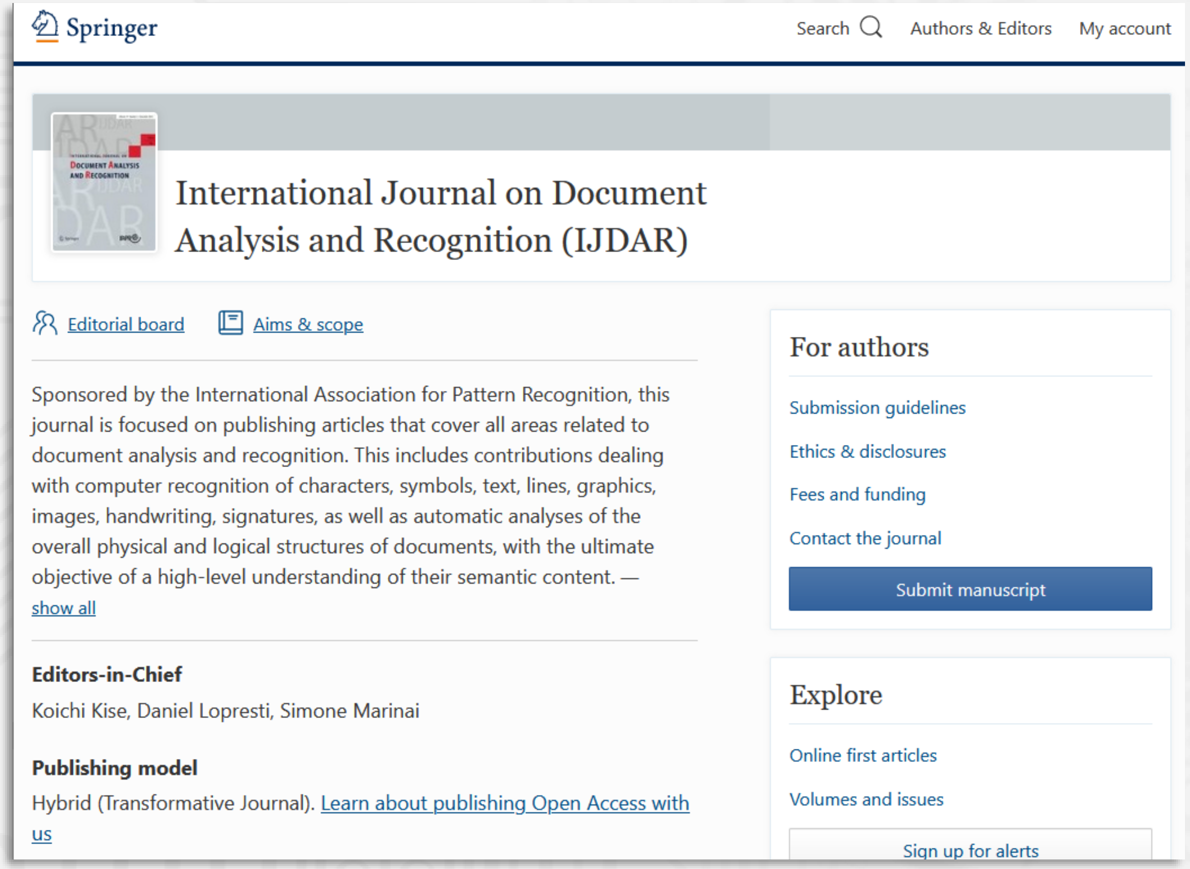
DAS 2020
Wuhan, China

14TH IAPR INTERNATIONAL WORKSHOP ON DOCUMENT ANALYSIS SYSTEMS

July 26~29, 2020, hosted from Wuhan, China

- Unlike other more general conferences in the field, DAS focuses on systems-related issues, although often related papers appear as well (e.g., classifier techniques).

Popular venue: IJDAR



International Journal on Document Analysis and Recognition (IJDAR)

Sponsored by the International Association for Pattern Recognition, this journal is focused on publishing articles that cover all areas related to document analysis and recognition. This includes contributions dealing with computer recognition of characters, symbols, text, lines, graphics, images, handwriting, signatures, as well as automatic analyses of the overall physical and logical structures of documents, with the ultimate objective of a high-level understanding of their semantic content. — show all

Editors-in-Chief

Koichi Kise, Daniel Lopresti, Simone Marinai

Publishing model

Hybrid (Transformative Journal). Learn about publishing Open Access with us

For authors

Explore

- International Journal on Document Analysis and Recognition (Springer) is devoted to the field.
- Topics include pen-based computing (signature verification) and document authentication / validation, among many others.

Some recent papers (1)

ICDAR 2019:

- Offline Writer Identification Based on the Path Signature Feature.
- GRK-Papyri: A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification.
- Online Writer Identification using GMM Based Feature Representation and Writer-Specific Weights.
- A Spatio-Spectral Hybrid Convolutional Architecture for Hyperspectral Document Authentication, M. J. Khan, K. Khurshid, F. Shafait.
- Deep Dynamic Time Warping: End-to-End Local Representation Learning for Online Signature Verification.
- Capturing Micro Deformations from Pooling Layers for Offline Signature Verification.
- Offline Signature Verification using Structural Dynamic Time Warping.
- Online Signature Verification by Few-Shot Separable Convolution Based Deep Learning.

Some recent papers (2)

DAS 2018:

- Encoding CNN Activations for Writer Recognition.
- Gaussian Process Classification as Metric Learning for Forensic Writer Identification.
- Stable Regions and Object Fill-Based Approach for Document Images Watermarking.
- Towards Detection of Morphed Face Images in Electronic Travel Documents, U. Scherhag, C. Rathgeb, and C. Busch.
- A New Descriptor for Pattern Matching: Application to Identity Document Verification.
- Offline Bengali Writer Verification by PDF-CNN and Siamese Net.
- Saliency-Based Detection of Identity Documents Captured by Smartphones.
- Automated Forgery Detection in Multispectral Document Images Using Fuzzy Clustering.

IJDAR March 2020:

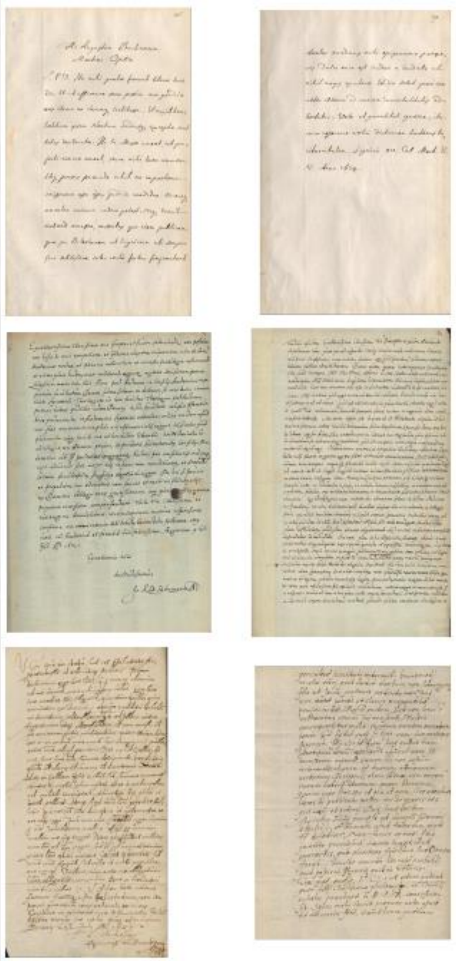
- Even big data is not enough: need for a novel reference modelling for forensic document authentication, U. Garain and B. Halder.

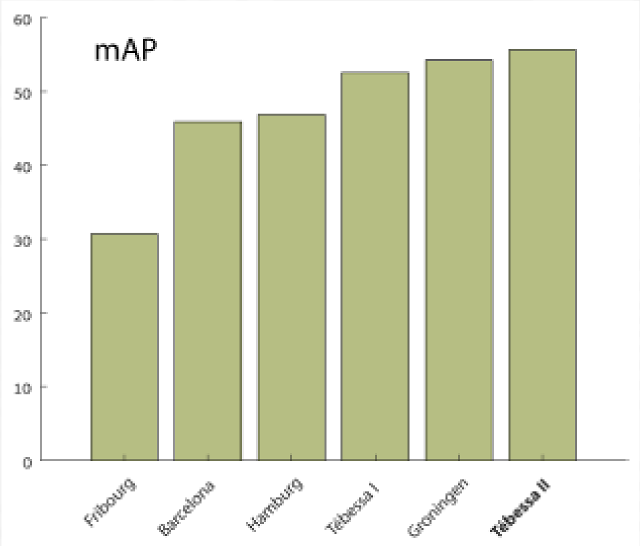
Competitions

- ICDAR 2021 Competition on On-Line Signature Verification.
- ICDAR 2019 Competition on Signature Verification based on an On-line and Off-line Signature Dataset.
- ICDAR 2017 Competition on Historical Document Writer Identification (Historical-WI).
- ICDAR 2017 Competition on Multi-script Writer Identification Using LAMIS-MSHD and CERUG Databases.
- ICDAR 2015 Signature Verification and Writer Identification Competitions for On- and Offline Skilled Forgeries (SigWIComp-2015).
- ICDAR 2015 Multi-Script Writer Identification and Gender Classification (MS-WIGC-2015).

ICDAR 2017 Historical-WI

- Goal was retrieval of pages which have been written by same author.
- Test dataset consisted of 3,600 handwritten pages originating from 13th to 20th Century.
- Manuscripts from 720 different writers where each writer contributed five pages.
- Five different institutions submitted six methods which were ranked using identification and retrieval metrics.





Some current problems of interest

As gleaned from upcoming ICDAR 2021 conference:

- Automating document layout analysis.
- Scene text recognition.
- Deep learning approaches for information extraction.
- Making historical manuscript images searchable.
- Extracting data from images of charts / tables / census reports.
- Authorship determination / writer identification.
- Asking questions on document collections / image text.

Some of my own work over the years (1)

- “Quantifying Information Leakage in Document Redaction,” D. Lopresti and A. L. Spitz, Proceedings of the First ACM Workshop on Hardcopy Document Processing (in association and Knowledge Management), November 2004, Washington, DC, p
- “Information Leakage Through Document Redaction: Attacks and C Spitz, Proceedings of Document Recognition and Retrieval XII (IS& Electronic Imaging), January 2005, San Jose, CA, pp. 183-190.

Resources can be combined in an interactive system to undo attempts to hide information via redaction in certain cases

- “The Effectiveness of Generative Attacks on an Online Handwriting Biometric,” D. Lopresti and J. Raim, Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication, July 2005, Rye Brook, NY, pp. 1090-1099.
- “Biometric Authentication Revisited: Understanding the Impact of Wo Monrose, and D. Lopresti, Proceedings of the Fifteenth USENIX Se Vancouver, BC, Canada, pp. 29-41.
- “Evaluating the Security of Handwriting Biometrics,” L. Ballard, D. L Tenth International Workshop on Frontiers in Handwriting Recogniti 461-466.

Traditional attack models for behavioral biometric security underestimate talented, resourceful adversaries

- “Forgery Quality and its Implications for Behavioral Biometric Security,” L. Ballard, D. Lopresti, and F. Monrose, IEEE Transactions on Systems, Man and Cybernetics, Part B, vol. 37, no. 5, October 2007, pp. 1107-1118.

Some of my own work over the years (2)

- “Biometric Key Generation Using Pseudo-Signatures,” L. Ballard, J. Chen, D. Lopresti, and F. Monrose, Proceedings of the Eleventh International Conference on Frontiers in Handwriting Recognition, August 2008, Montréal, Canada, pp. 646-651.
- “Pseudo-Signatures as a Biometric,” J. Chen, D. Lopresti, L. Ballard Second IEEE International Conference on Biometrics: Theory, Appl 2008, Arlington, VA, pages 6 (CD-ROM).

Pseudo-signatures (graphical passwords) can overcome some inherent issues with using real signatures

- “Toward Resisting Forgery Attacks via Pseudo-Signatures,” J. Chen, D. Lopresti, and F. Monrose, Proceedings of the Tenth International Conference on Document Analysis and Recognition, July 2009, Barcelona, Spain, pp. 51-55.
- “The Impact of Ruling Lines on Writer Identification,” J. Chen, D. Lopresti, and E. Kavallieratou, Proceedings of the Twelfth International Conference on Frontiers in Handwriting Recognition, November 2010, Kolkata, India, pp. 439-444.
- “Using Perturbed Handwriting to Support Writer Identification in the Chen, W. Cheng, and D. Lopresti, Document Recognition and Retri Symposium on Electronic Imaging), January 2011, San Francisco, C

Methods for improving performance of writer identification in face of severe constraints on training data

- “Parameter Calibration for Synthesizing Realistic-Looking Variability in Offline Handwriting,” W. Cheng and D. Lopresti, Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging), January 2011, San Francisco, CA, pp. 78740Y-1 - 78740Y-10.

Some of my own work over the years (3)

- "A New Method For Detecting Altered Text in Document Images," L. D. Lopresti, B. Seraogi, and B. B. Chaudhuri, Proceedings of the Se Recognition and Artificial Intelligence (ICPRAI 2020), October 2020
- "Forged Text Detection in Video, Scene, and Document Images," L. S. Raghunandan, U. Pal, T. Lu, and D. Lopresti, to appear in IET Im

Deep learning combined with targeted image processing steps to detect copy-paste and insertion forgeries

- "Camera-based Ballot Counter," G. Nagy, B. Clifford, A. Berg, G. Saunders, D. Lopresti, and E. Barney Smith, Proceedings of the Tenth International Conference on Document Analysis and Recognition, July 2009, Barcelona, Spain, pp. 151-155.
- "Style-Based Ballot Mark Recognition," P. Xiu, D. Lopresti, H. Baird, of the Tenth International Conference on Document Analysis and Re 216-220.
- "Document Analysis Issues in Reading Optical Scan Ballots," D. Lo Proceedings of the Ninth IAPR International Workshop on Documen pp. 105-112.

Approaches to facilitate the trustworthy capture and reading of election ballots

- "Characterizing Challenged Minnesota Ballots," G. Nagy, D. Lopresti, E. H. Barney Smith, and Z. Wu, Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging), January 2011, San Francisco, CA, pp. 787413-1 - 787413-10.

Other useful resources: IAPR

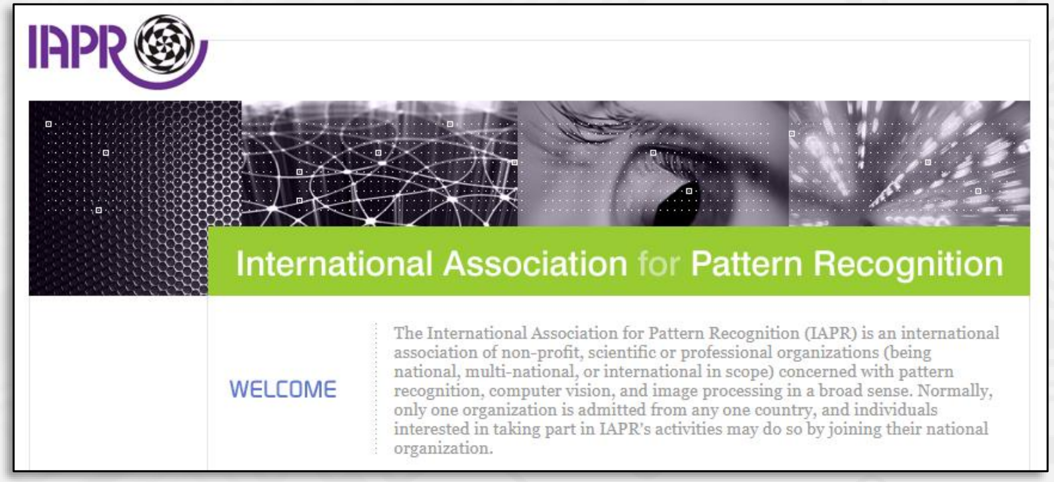
IAPR – the International Association for Pattern Recognition – is the organization responsible for supporting a wide range of activities by the research community.

- IAPR organizes conferences, administers awards.

IAPR technical committees
do much of the actual work:

- TC-06 (Computational Forensics),
- TC-10 (Graphics Recognition),
- TC-11 (Reading Systems).

See www.iapr.org.

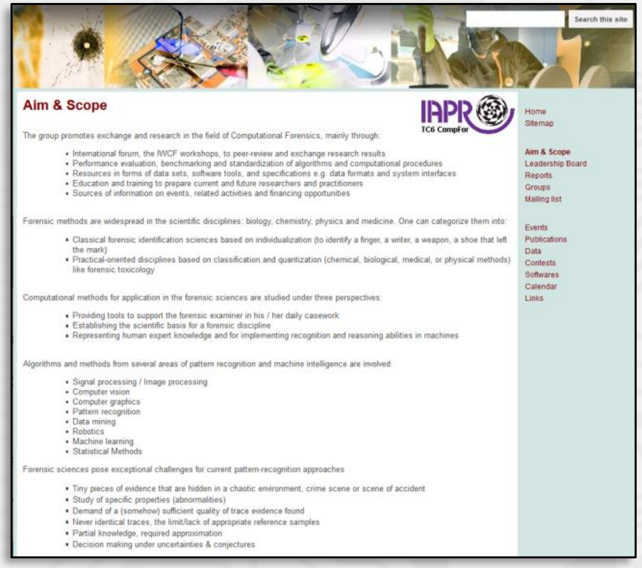


The International Association for Pattern Recognition (IAPR) is an international association of non-profit, scientific or professional organizations (being national, multi-national, or international in scope) concerned with pattern recognition, computer vision, and image processing in a broad sense. Normally, only one organization is admitted from any one country, and individuals interested in taking part in IAPR's activities may do so by joining their national organization.

Other useful resources: TC-06

IAPR TC-06 (“Computational Forensics”) aims and scope:

- International forum, the IWCF workshops, to peer-review and exchange research results.
- Performance evaluation, benchmarking and standardization of algorithms and computational procedures.
- Resources in forms of data sets, software tools, and specifications e.g. data formats and system interfaces.
- Education and training to prepare current and future researchers and practitioners.
- Sources of information on events, related activities and financing opportunities.



Aim & Scope

The group promotes exchange and research in the field of Computational Forensics, mainly through:

- International forum, the IWCF workshops, to peer-review and exchange research results
- Performance evaluation, benchmarking and standardization of algorithms and computational procedures
- Resources in forms of data sets, software tools, and specifications e.g. data formats and system interfaces
- Education and training to prepare current and future researchers and practitioners
- Sources of information on events, related activities and financing opportunities

Forensic methods are widespread in the scientific disciplines: biology, chemistry, physics and medicine. One can categorize them into:

- Classical forensic identification sciences based on individualization (to identify a finger, a writer, a weapon, a shoe that left the mark)
- Practical-oriented disciplines based on classification and quantization (chemical, biological, medical, or physical methods) like forensic toxicology

Computational methods for application in the forensic sciences are studied under three perspectives:

- Providing tools to support the forensic examiner in his / her daily casework
- Establishing the scientific basis for a forensic discipline
- Representing human expert knowledge and for implementing recognition and reasoning abilities in machines

Algorithms and methods from several areas of pattern recognition and machine intelligence are involved:

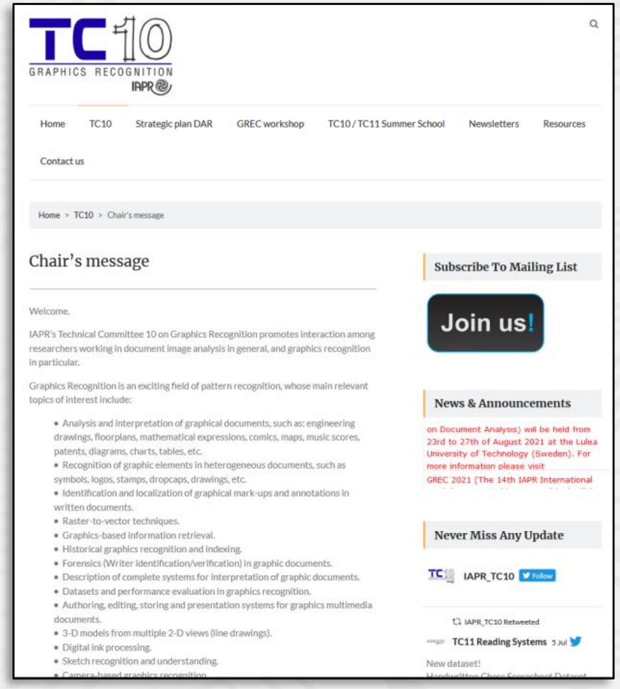
- Signal processing / Image processing
- Computer vision
- Computer graphics
- Pattern recognition
- Data mining
- Robotics
- Machine learning
- Statistical Methods

Forensic sciences pose exceptional challenges for current pattern-recognition approaches

- Tiny pieces of evidence that are hidden in a chaotic environment, crime scene or scene of accident
- Study of specific properties (abnormalities)
- Demand of a (somehow) sufficient quality of trace evidence found
- Never identical traces, the limit/lack of appropriate reference samples
- Partial knowledge, required approximation
- Decision making under uncertainties & conjectures

Other useful resources: TC-10

TC-10 (“Graphics Recognition”) promotes interaction among researchers working in document image analysis in general, and graphics recognition in particular.



Chair’s message

Welcome.

IAPR’s Technical Committee 10 on Graphics Recognition promotes interaction among researchers working in document image analysis in general, and graphics recognition in particular.

Graphics Recognition is an exciting field of pattern recognition, whose main relevant topics of interest include:

- Analysis and interpretation of graphical documents, such as: engineering drawings, floorplans, mathematical expressions, comics, maps, music scores, patents, diagrams, charts, tables, etc.
- Recognition of graphic elements in heterogeneous documents, such as symbols, logos, stamps, dropcaps, drawings, etc.
- Identification and localization of graphical mark-ups and annotations in written documents.
- Raster-to-vector techniques.
- Graphics-based information retrieval.
- Historical graphics recognition and indexing.
- Forensics (Writer identification/verification) in graphic documents.
- Description of complete systems for interpretation of graphic documents.
- Datasets and performance evaluation in graphics recognition.
- Authoring, editing, storing and presentation systems for graphics multimedia documents.
- 3-D models from multiple 2-D views (line drawings).
- Digital ink processing.
- Sketch recognition and understanding.
- Camera-based graphics recognition

Subscribe To Mailing List

News & Announcements

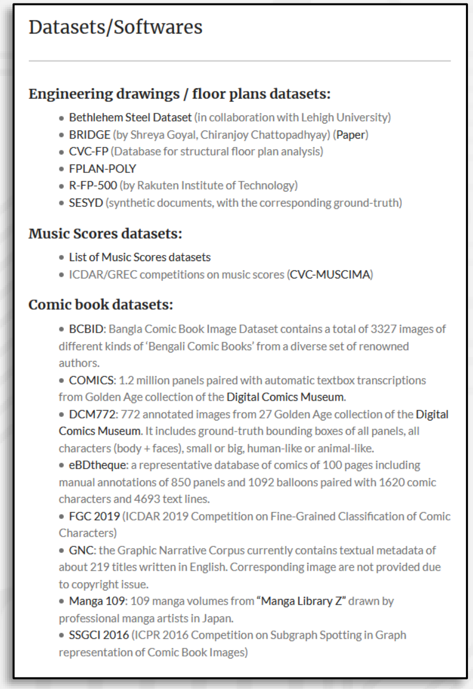
on Document Analysis) will be held from 23rd to 27th of August 2021 at the Lulea University of Technology (Sweden). For more information please visit

GREC 2021 (The 14th IAPR International

Never Miss Any Update

New dataset!

Handwritten Chess Scoresheet Dataset



Datasets/Softwares

Engineering drawings / floor plans datasets:

- Bethlehem Steel Dataset (in collaboration with Lehigh University)
- BRIDGE (by Shreya Goyal, Chiranjoy Chattopadhyay) (Paper)
- CVC-FP (Database for structural floor plan analysis)
- FPLAN-POLY
- R-FP-500 (by Rakuten Institute of Technology)
- SESYD (synthetic documents, with the corresponding ground-truth)

Music Scores datasets:

- List of Music Scores datasets
- ICDAR/GREC competitions on music scores (CVC-MUSCIMA)

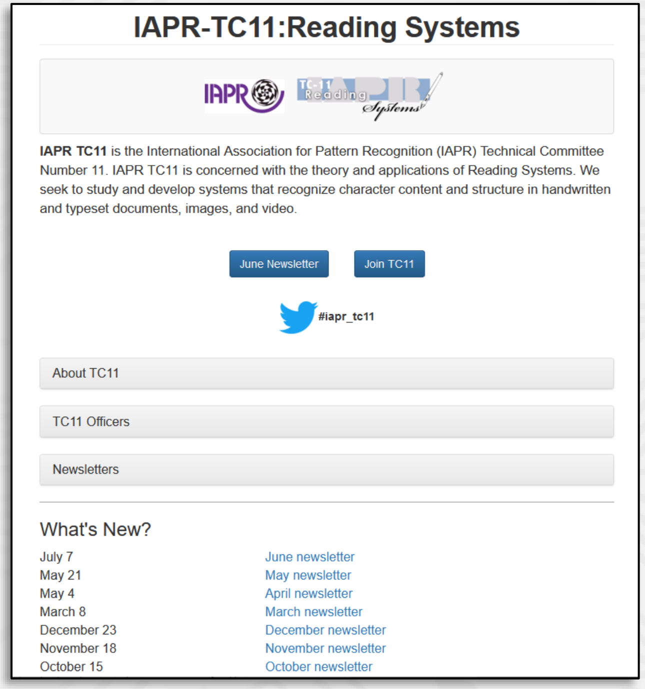
Comic book datasets:

- BCBID: Bangla Comic Book Image Dataset contains a total of 3327 images of different kinds of ‘Bengali Comic Books’ from a diverse set of renowned authors.
- COMICS: 1.2 million panels paired with automatic textbox transcriptions from Golden Age collection of the Digital Comics Museum.
- DCM772: 772 annotated images from 27 Golden Age collection of the Digital Comics Museum. It includes ground-truth bounding boxes of all panels, all characters (body + faces), small or big, human-like or animal-like.
- eBDtheque: a representative database of comics of 100 pages including manual annotations of 850 panels and 1092 balloons paired with 1620 comic characters and 4693 text lines.
- FGC 2019 (ICDAR 2019 Competition on Fine-Grained Classification of Comic Characters)
- GNC: the Graphic Narrative Corpus currently contains textual metadata of about 219 titles written in English. Corresponding image are not provided due to copyright issue.
- Manga 109: 109 manga volumes from “Manga Library Z” drawn by professional manga artists in Japan.
- SSGCI 2016 (ICPR 2016 Competition on Subgraph Spotting in Graph representation of Comic Book Images)

Available datasets including engineering drawings, musical scores, very large sets of comic book images.

Other useful resources: TC-11

TC-11 (“Reading Systems”) represents the international research community in topics relating to character recognition and document analysis.

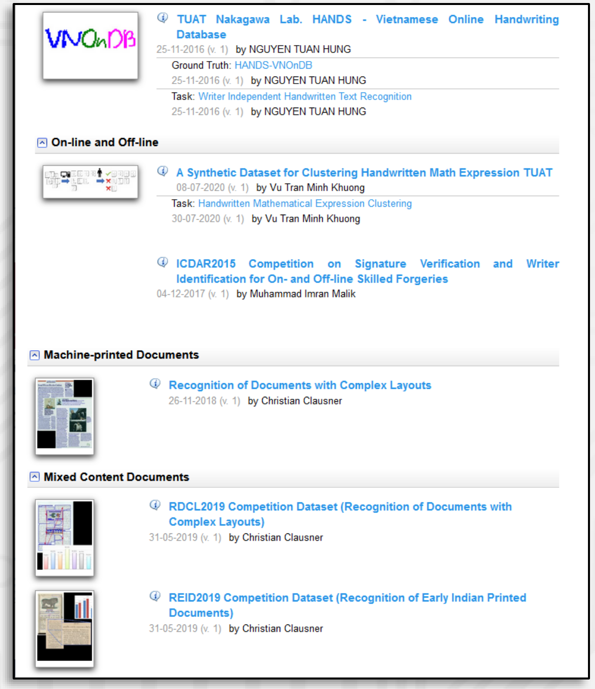


IAPR-TC11:Reading Systems

IAPR TC11 is the International Association for Pattern Recognition (IAPR) Technical Committee Number 11. IAPR TC11 is concerned with the theory and applications of Reading Systems. We seek to study and develop systems that recognize character content and structure in handwritten and typeset documents, images, and video.

What's New?

July 7	June newsletter
May 21	May newsletter
May 4	April newsletter
March 8	March newsletter
December 23	December newsletter
November 18	November newsletter
October 15	October newsletter



TUAT Nakagawa Lab. HANDS - Vietnamese Online Handwriting Database
25-11-2016 (v. 1) by NGUYEN TUAN HUNG
Ground Truth: HANDS-VNOnDB
25-11-2016 (v. 1) by NGUYEN TUAN HUNG
Task: Writer Independent Handwritten Text Recognition
25-11-2016 (v. 1) by NGUYEN TUAN HUNG

On-line and Off-line

A Synthetic Dataset for Clustering Handwritten Math Expression TUAT
08-07-2020 (v. 1) by Vu Tran Minh Khuong
Task: Handwritten Mathematical Expression Clustering
30-07-2020 (v. 1) by Vu Tran Minh Khuong

ICDAR2015 Competition on Signature Verification and Writer Identification for On- and Off-line Skilled Forgeries
04-12-2017 (v. 1) by Muhammad Imran Malik

Machine-printed Documents

Recognition of Documents with Complex Layouts
26-11-2018 (v. 1) by Christian Clausner

Mixed Content Documents

RDCL2019 Competition Dataset (Recognition of Documents with Complex Layouts)
31-05-2019 (v. 1) by Christian Clausner

REID2019 Competition Dataset (Recognition of Early Indian Printed Documents)
31-05-2019 (v. 1) by Christian Clausner

Many available datasets, including signature verification, writer identification, multispectral images of ancient documents.

Observations and synergies

- Shared interests: shared methods and shared applications.
- Valuable things to be learned looking in both directions.
- Differences in research culture? But these can be bridged.
- Not as much “cross-fertilization” as there could be.

Building connections:

- Sign up for IAPR TC-06 and/or TC-11 mailing lists.
- Peruse resources highlighted in this talk.
- Submit papers to ICDAR (or affiliated workshops), DAS, IJDAR.
- Propose a new, complementary workshop or competition.

Looking forward

- DAS 2022 will be held in La Rochelle, France in May 2022.

ICDAR 2023

The 17th International Conference on Document Analysis and Recognition

August, 2023 — San Jose, California

- ICDAR 2023 will be held in San Jose, CA in August 2023.
- Calls for Papers, Competitions, and Workshop proposals should be going out soon.

Thank you

I look forward to the day in the future when we can gather together once again for professional meetings like this.

In the meantime, please feel free to contact me with any questions / comments / suggestions:

lopresti@cse.lehigh.edu

http://www.cse.lehigh.edu/~lopresti/