Machine Vision

Lecture 13
Document Image Analysis

Based on lectures of

Henry S. Baird

Palo Alto Research Center

parc
Palo Alto Research Center

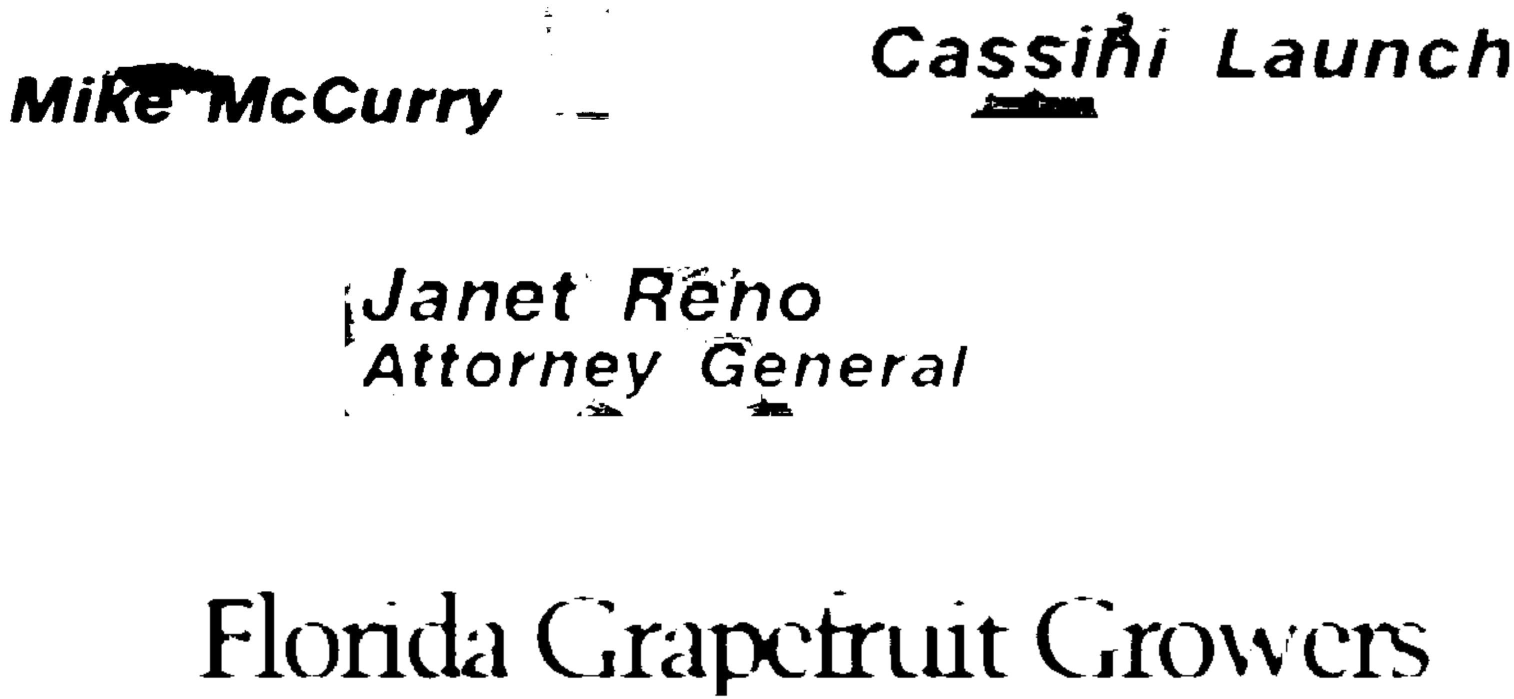
A walking tour of the Document Image Analysis research field

- Machine ‘reading’ of text, maps, music scores, ...
- History & kinship to Computer Vision
- Pressing open problems
- Digital Libraries
- Web Security

A Classic Problem Instance

- Given a digital image of a document (TIFF, PNB, …)
- Separate **text** from **non-text** (photos, graphics, ...)
- Locate **columns** of text
- … **lines** of text
- … **words**
- … **characters**
- Recognize the text
- Label parts by function (title, author, …)
- Output text in encoded form (ASCII, XML, UNICODE, …)

Examples illustrating problems



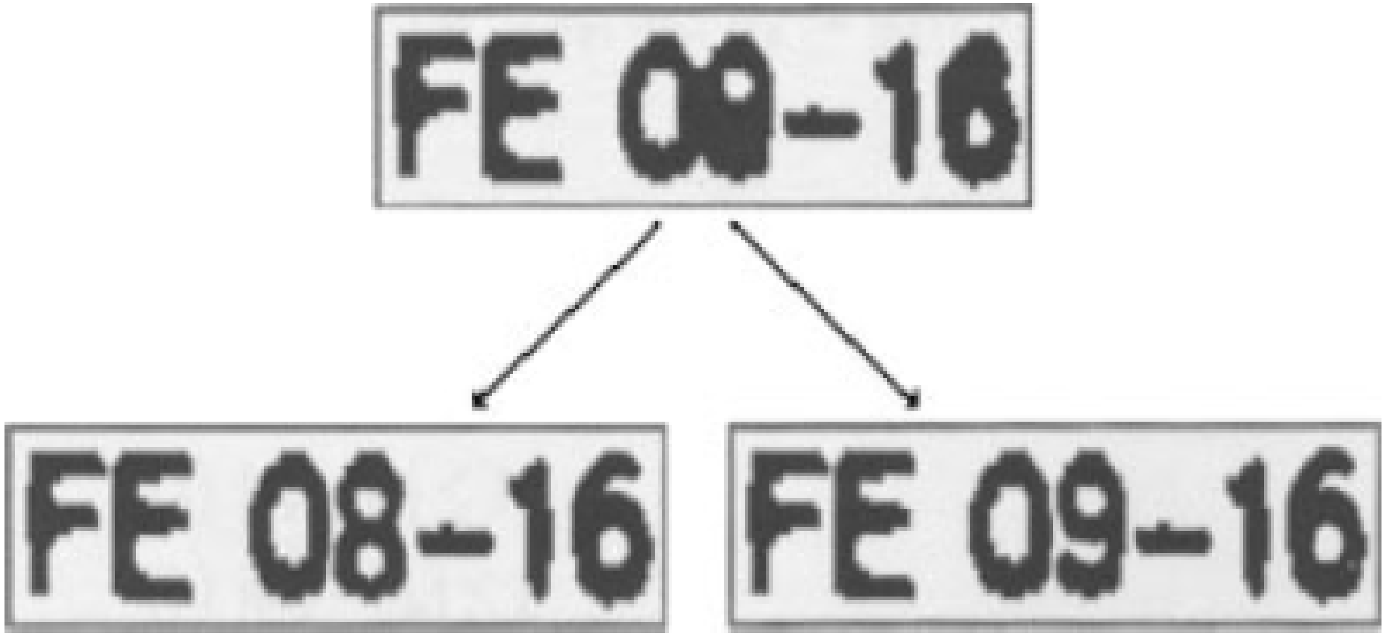


Figure 67 Ambiguity in character recognition [114]

Example of doorplate to be recognized by robot

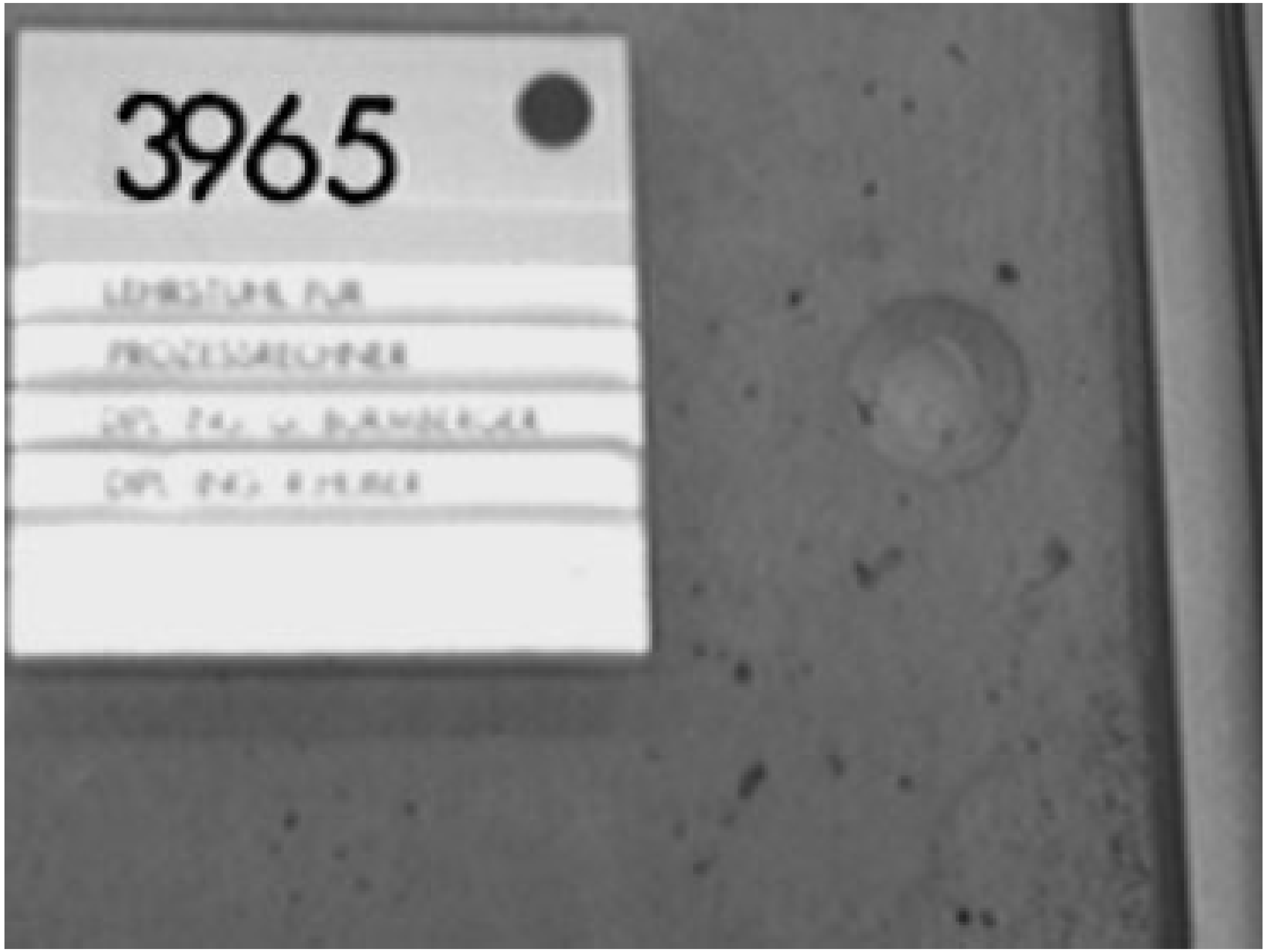


Figure 68 Characters that are stuck together

Example of doorplate to be recognized by robot

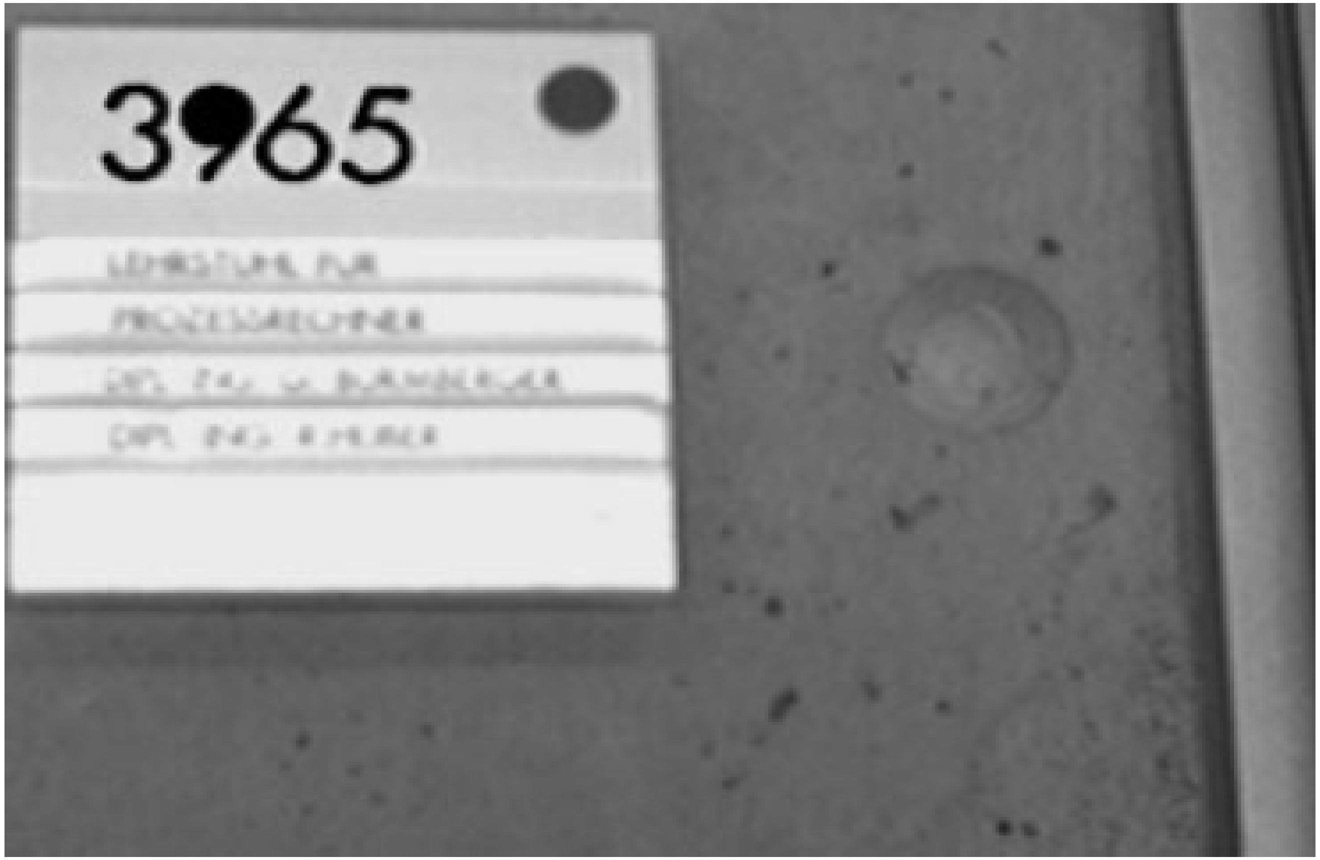


Figure 69 Merging within a character

Example of doorplate to be recognized by robot

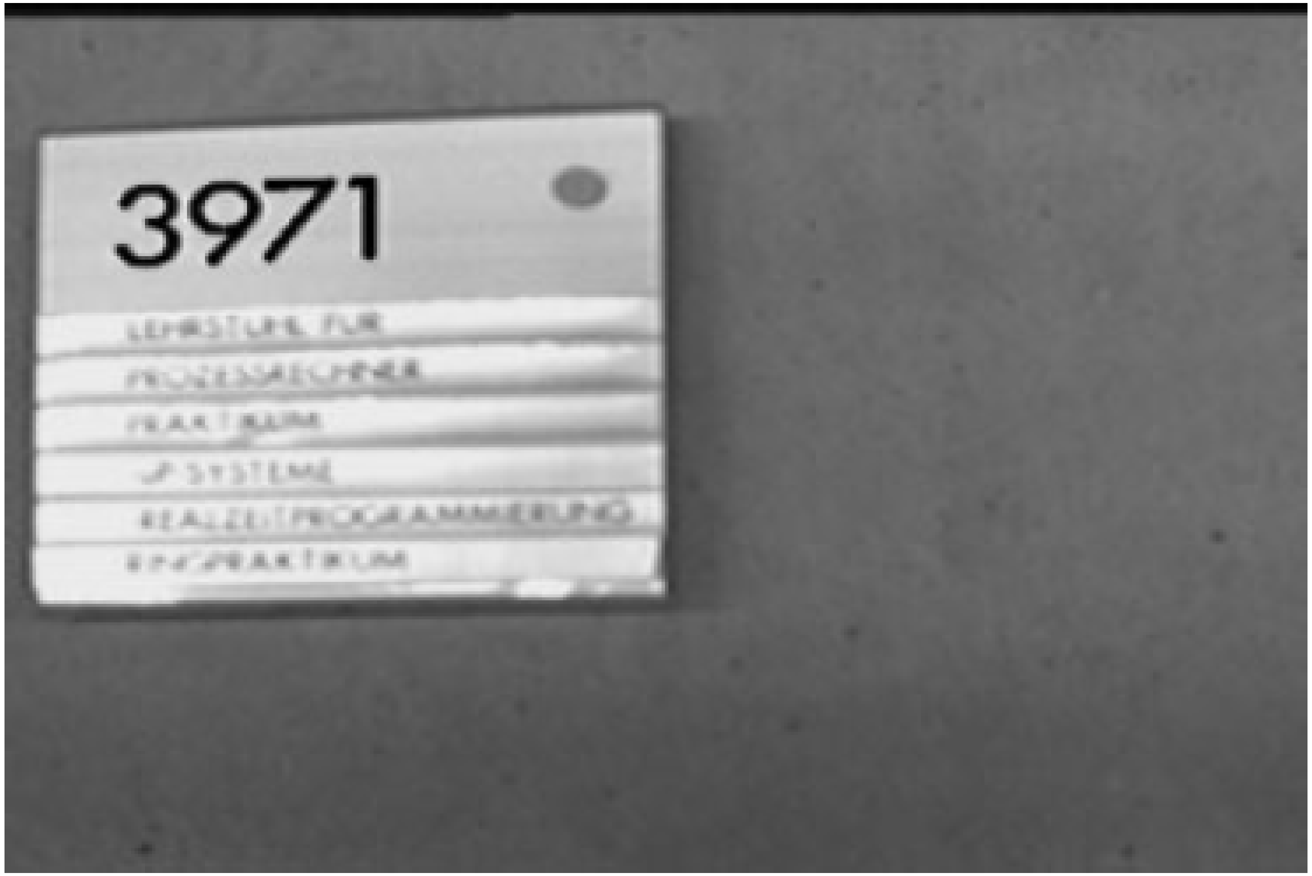


Figure 70 Similar numerals

Example of doorplate to be recognized by robot

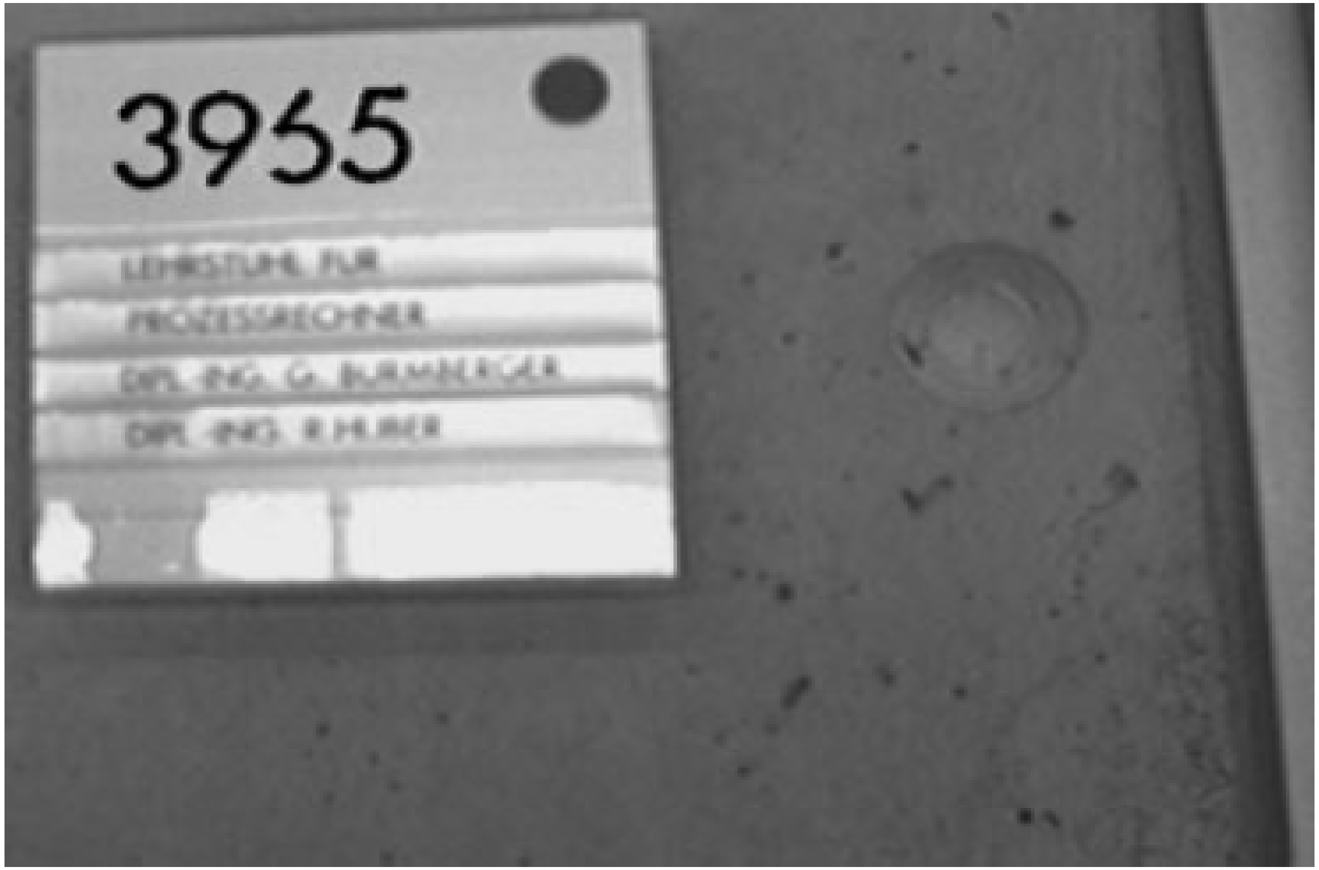


Figure 71 A numeral that is not closed

Most Familiar Product: Desk-top OCR Software

- OCR = Optical Character Recognition
- Small office, home office (SOHO)
  - Casual-use “page readers”
  - fully automatic, but unimprovable
- Mature international market
  - sometimes near-perfect; often wretched
  - ScanSoft,Toshiba, Abbyy, Tsinghua, …
  - no clear performance leader on English:
    - » commodity pricing
- Steady but slow progress:
  - 15-25% fewer errors / year

Text, and more ...

- **Text:**

  running text, addresses, checks, tables, ...
- **Graphics:**

  forms, maps, drawings, architectural plans, ...
- **Special notations:**

  music, mathematics, chemical diagrams, ...
- Machine-printed *vs* handwritten
- Off-line (static) *vs* on-line (dynamic)

Affinities with Computer Vision

- Signal $\rightarrow$ Symbol
  - can't directly measure what we want
  - noisy, underdetermined problems
- Ambitious goals:
  - accurate interpretation of complex images
- Explicit 'priors' (models) are crucial
  - to supply the implicit context
- Complete models rarely available
  - can weak models succeed?
  - can strong models be trained or inferred?

Weak Models:
e.g., Postal Addresses

- Largely unconstrained:

  typefaces, writing styles, ink color

- Variable layouts, background shading, ...

- However, known constraints on:
  - city/state/ZIP combinations
  - **hugely** helpful

Postal Address Reading

- Largest application worldwide:
  - USA, France, Germany, Japan, …
- Huge economic impact:
  - offsets **7 cents** of US postage
- Still far from perfect:
  - ~35% of HW addresses rejected

Strong Models: e.g., Barcodes etc

- Controlled pattern, size, ink, light, scanner
- Error-correcting codes
- Orders of magnitude more accurate, fast
  - error rates in parts / million
- Confined to niche markets

Checks, etc

- Rapid adoption of check readers by banks
  - off-line, handwritten and machine-printed
- Combining evidence:
  - e.g., from courtesy and legal amounts
- Varied check layouts a challenge:
  - US business checks nonstandardized
- Background clutter a serious problem:
  - US personal checks “individualized”

Graphics Recognition

- Line-graphics + text
  - **fixed forms**: a mature field
  - **maps**: some early successes
  - **engineering drawings**: exploratory
  - **chemical diagrams**: exploratory
- A strongly developing subfield of DIA
  - GREC workshop, DBs, competitions
- Key technical challenges:
  - extraction of primitives: lines, arcs, etc
  - flexible geometric models
  - integration of evidence across ‘levels’:
    - » primitives, shapes, connections, semantics

Music OCR

- International, “language-free” problem
- Difficult physical segmentation:
  - overprinting, stretchable symbols
- Rich domain for systems exploration:
  - exploiting domain-specific knowledge
  - control flow and optimization
- Research $\rightarrow$ Products in 10 years

Expanding Research Domain

A. M. Turing’s plan: reader for the blind

50s: machine-print fixed font & size

60s: fixed forms, OCR-A/B fonts

70s: multi-font, variable size, handwriting

80s: variable layouts, language context

90s: multi-lingual, graphics, tables, music, math

2000+: digital libraries, paper/digital portals

Now a Distinct R&D Community

- Through mid-1980's, DIA was part of
  "early AI" = AI+PR+IP+CV
  - conferences: ICPR, CVPR
  - journals: PAMI, PR, PRL
- Then, a wave of specialization split it up…
- In 1990's, DIA came into its own:
  - conferences: ICDAR, SDAIR, DR&R
  - workshops: DAS, IWFHR, GREC, DLIA, WDA
  - journal: IJDAR

Support for Research in USA

- Postal Services $$$$$
- DARPA/DOD $$$
- DOE $$
- NSF $ (DLI)
- Desktop OCR   hire PhDs
- Banking/Finance   buy products

DIA has Evolved a Little Differently from CV

- Cultural, not physical, context (mostly):
  - **input:** messages -- *not* natural scenes
  - **goal**: assist communication -- *not* make artificial HVS
  - **models:** intention, meaning, language, alphabets,
    glyphs, layout, printing, scanning, …
    -- *not* physics of light, motion, …
- Consensus on methodology:
  - performance metrics
  - large-scale empirical evaluation
- Close association with engineers & users:
  - established, growing commercial niches
  - systems engineering is a DIA *research* area

Most Closely Allied Disciplines

- **Computer Vision**
- Pattern Recognition / Decision Theory
- Statistics / Machine Learning
- Information Retrieval
- Computational Linguistics
- Computational Geometry *layout analysis*
- Speech Recognition *HMMs, transducers*
- Psychophysics (of reading)
- Digital Libraries
- Human Interactive Proofs

Technical Challenges
… in Text Recognition

- Symbol sets: 30-30,000
- Typefaces: 1000s
- Language and other context
- Page layouts
- Image quality

DIA R&D for Image Quality Control

- Measuring document image quality
  - new test target designs
  - image processing algorithms
  - rigorous, quantitative standards
- Assuring quality
  - fast algorithms for on-the-fly image quality estimation
- Predicting human & machine legibility
  - What image quality features correlate
  - well with human and OCR legibility?
  - … and with other, later DIA tasks?

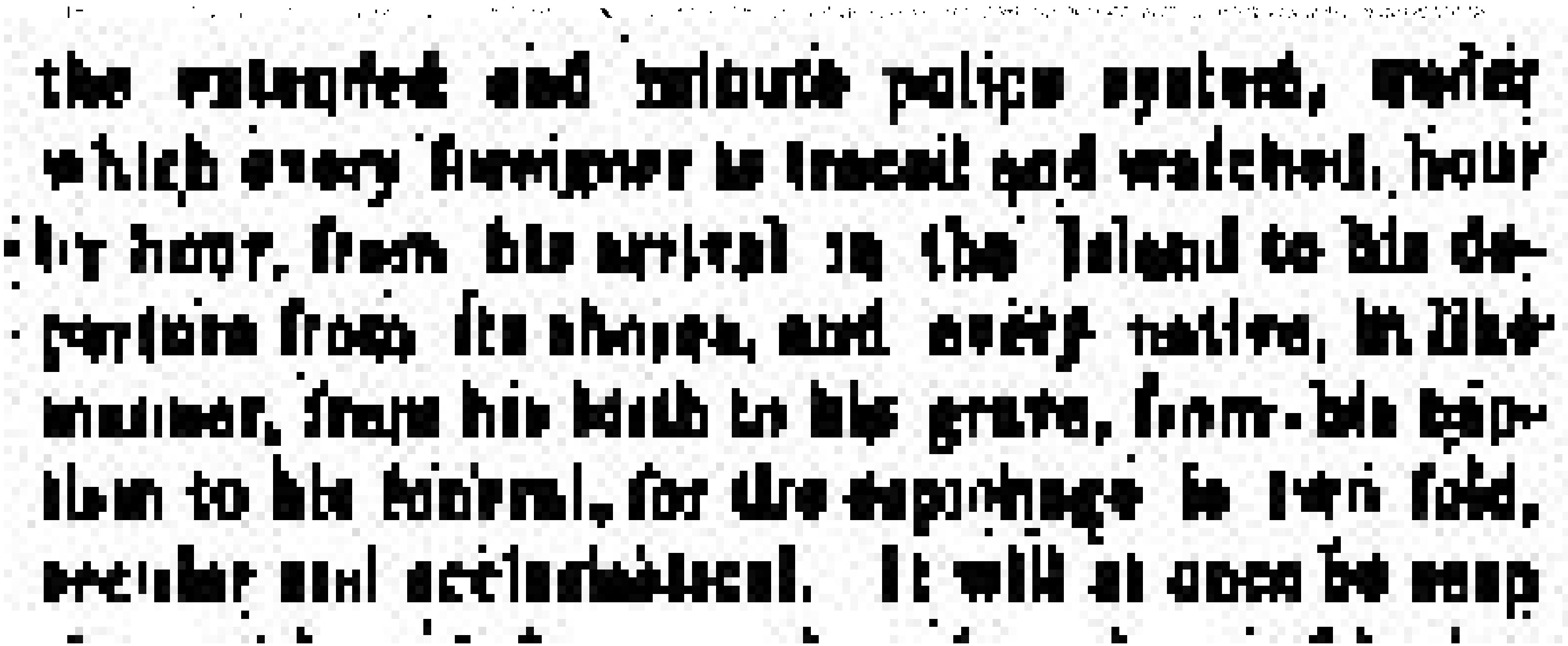
K. Summers, "Document Image Improvement for OCR as a Classification Problem," *Proc., DR&R X*, Santa Clara,CA, Jan 2003.

E. H. Barney Smith & X. Qiu, "Relating Statistical Image Differences & Degradation Features," *Proc*, *5th DAS*, Princeton, NJ., Aug 2002.

When Quality Control Goes Wrong

Front Page, 1852 Edition of the New York Times

Scanned from microfilm.



The Historical New York Times Project, CMU/NYT, 1999.

Extracting & Recognizing Content

These are central DIA R&D goals

But existing doc image understanding systems

cannot guarantee high accuracy

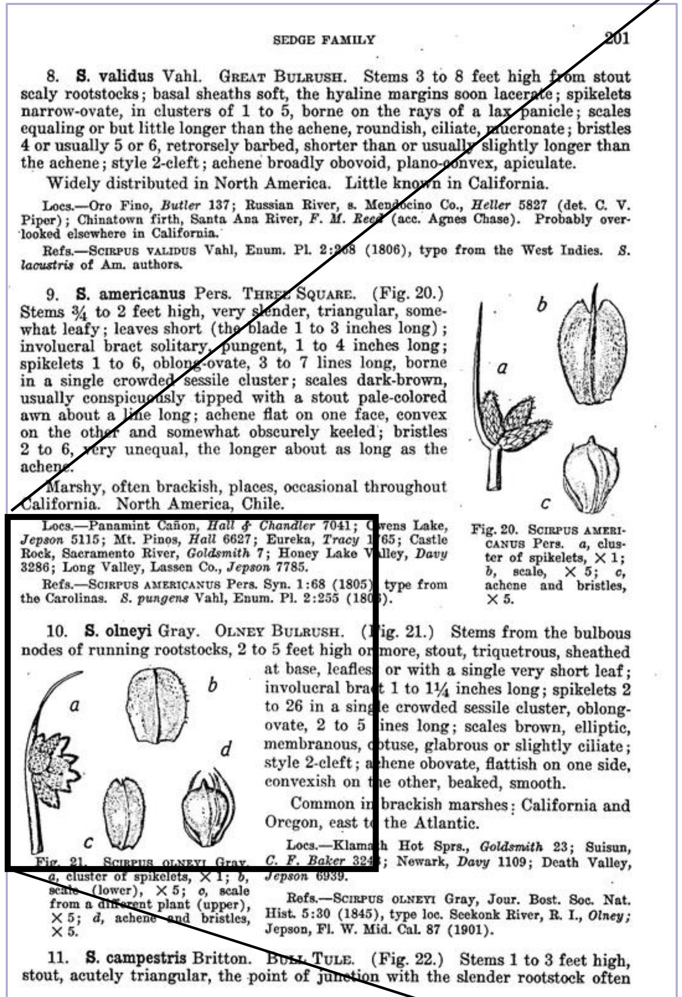
across the full range of documents:

» - typefaces, h/w styles	old fashioned
» - image qualities	poor & variable
» - layout geometries	deformed
» - writing systems	obsolete
» - languages	rare
» - domains of discourse	arcane

DL's scholarly & historical docs are often harder

S. Rice, G. Nagy, T. Nartker, *OCR: An Illustrated Guide to the Frontier*,
Kluwer Academic Publishers: 1999.

Richly Meaningful Typographical Book Designs



8. **S. validus** Vahl. GREAT BULRUSH. Stems 3 to 8 feet high from stout scaly rootstocks; basal sheaths soft, the hyaline margins soon lacerate; spikelets narrow-ovate, in clusters of 1 to 5, borne on the rays of a lax panicle; scales equaling or but little longer than the achene, roundish, ciliate, mucronate; bristles 4 or usually 5 or 6, retrorsely barbed, shorter than or usually slightly longer than the achene; style 2-cleft; achene broadly obovoid, plano-convex, apiculate.

Widely distributed in North America. Little known in California.

Locs.—Oro Fino, *Butler* 137; Russian River, s. Mendocino Co., *Heller* 5827 (det. C. V. Piper); Chinatown firth, Santa Ana River, *F. M. Reed* (acc. Agnes Chase). Probably overlooked elsewhere in California.

Refs.—SCIRPUS VALIDUS Vahl, Enum. Pl. 2:268 (1806), type from the West Indies. *S. lacustris* of Am. authors.

9. **S. americanus** Pers. THREE SQUARE. (Fig. 20.) Stems ¾ to 2 feet high, very slender, triangular, somewhat leafy; leaves short (the blade 1 to 3 inches long); involucral bract solitary, pungent, 1 to 4 inches long; spikelets 1 to 6, oblong-ovate, 3 to 7 lines long, borne in a single crowded sessile cluster; scales dark-brown, usually conspicuously tipped with a stout pale-colored awn about a line long; achene flat on one face, convex on the other and somewhat obscurely keeled; bristles 2 to 6, very unequal, the longer about as long as the achene.

Marshy, often brackish, places, occasional throughout California. North America, Chile.

Locs.—Panamint Cañon, *Hall & Chandler* 7041; Owens Lake, *Jepson* 5115; Mt. Pinos, *Hall* 6627; Eureka, *Tracy* 1765; Castle Rock, Sacramento River, *Goldsmith* 7; Honey Lake Valley, *Davy* 3286; Long Valley, Lassen Co., *Jepson* 7785.

Refs.—SCIRPUS AMERICANUS Pers. Syn. 1:68 (1805), type from the Carolinas. *S. pungens* Vahl, Enum. Pl. 2:255 (1806).

Fig. 20. SCIRPUS AMERICANUS Pers. *a*, cluster of spikelets, × 1; *b*, scale, × 5; *c*, achene and bristles, × 5.

10. **S. olneyi** Gray. OLNEY BULRUSH. (Fig. 21.) Stems from the bulbous nodes of running rootstocks, 2 to 5 feet high or more, stout, triquetrous, sheathed at base, leafless or with a single very short leaf; involucral bract 1 to 1¼ inches long; spikelets 2 to 26 in a single crowded sessile cluster, oblong-ovate, 2 to 5 lines long; scales brown, elliptic, membranous, obtuse, glabrous or slightly ciliate; style 2-cleft; achene obovate, flattish on one side, convexish on the other, beaked, smooth.

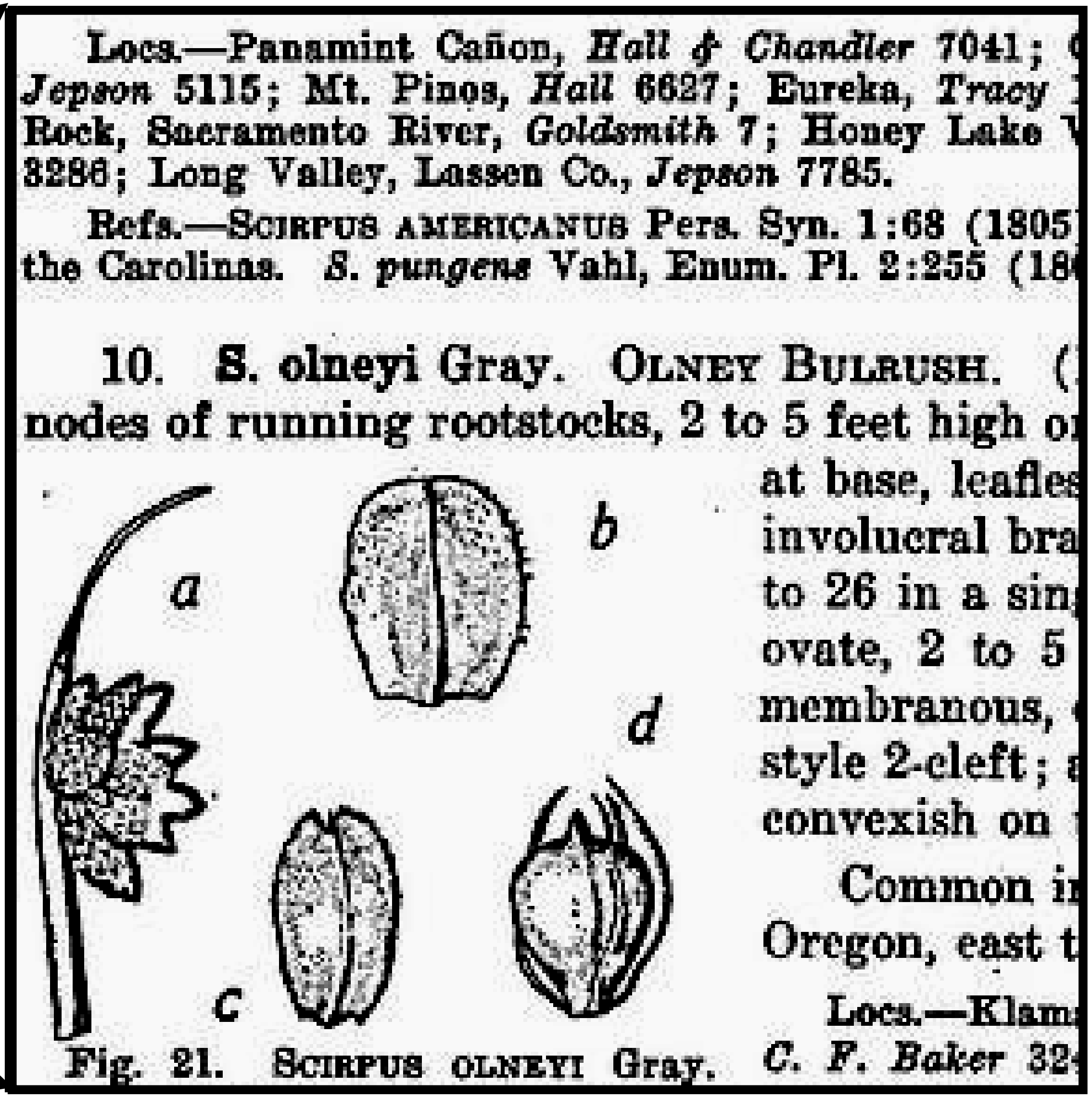
Common in brackish marshes: California and Oregon, east to the Atlantic.

Locs.—Klamath Hot Sprs., *Goldsmith* 23; Suisun, *C. F. Baker* 3243; Newark, *Davy* 1109; Death Valley, *Jepson* 6939.

Refs.—SCIRPUS OLNEYI Gray, Jour. Bost. Soc. Nat. Hist. 5:30 (1845), type loc. Seekonk River, R. I., *Olney*; Jepson, Fl. W. Mid. Cal. 87 (1901).

Fig. 21. SCIRPUS OLNEYI Gray. *a*, cluster of spikelets, × 1; *b*, scale (lower), × 5; *c*, scale from a different plant (upper), × 5; *d*, achene and bristles, × 5.

11. **S. campestris** Britton. BULL TULE. (Fig. 22.) Stems 1 to 3 feet high, stout, acutely triangular, the point of junction with the slender rootstock often



Make Doc-Images Highly Portable, Legible Everywhere

- No OCR errors!
- (Only layout errors.)
- Preserve meaningful
- appearance

- Challenges:
- reading order
- non-text
- navigation
- linking

Recognition, and more...

- Recognition
- Segmentation: parts of document
- Compression / coding
- Indexing & Retrieval
- Summarization
- Duplicate detection

Recognition / Segmentation / Compression

- Interrelated theoretically & practically:
  - perfect recognition is an ideal coding
  - segmentation assists recognition & coding
  - compression enables recognition
- Attacked piecemeal today
  - e.g., which to attempt first?
- Can they be simultaneously optimized?

Empirical Evaluation

- Early and lasting agreement w/in DIA field:
  - consensus on performance metrics
  - collect sample-image DBs w/ “ground truth”
  - extremely large-scale systematic testing
- Positive effects:
  - track industry-wide progress
  - raise the bar for publication (esp. journals)
  - identify the most pressing open problems
    - often surprising

Surprises So Far ...

- No Best Classifier
  - voting multiple-classifiers always dominate
- The Best Training Set Wins
  - size & representativeness is all
- Image Quality is Critical, but Imponderable
  - explains much failure, but hard to model
- Humans May be Beatable
  - The Bayes risk of concrete problems oddly low

Even in viable applications, Performance is Often Poor

- Many users remain badly served:
  - 40% **MP magazine pages**: *3-15% char error*
  - 37-55% **HW checks:** *rejected @ 1% error*
  - 35% **HW postal addresses:** *not ‘finalized’*
- Obstacles to progress:
  - systems too complex & unprincipled
  - riddled with special cases

Systems Architecture Research

- Embraced & encouraged
  - Systems papers are archivally publishable
  - Document Analysis Systems workshop series
  - DOE-, DOD-sponsored competitions
- Systems-architecture issues
  - design of versatile systems:
    - trainable, retargetable, adaptive
  - improving systems performance
    - error management, optimization

Accuracy / Versatility / Automation

Achieving *all three simultaneously*

is desirable, but elusive ….

- Sacrifice some accuracy:
  - desk-top OCR, IR *general-purpose, automatic*
- Sacrifice some versatility:
  - bar-codes *highly accurate, automatic*
- Sacrifice some automation:
  - table-readers, legacy conversion

Versatility is Particularly Hard

- “Polyfont” OCR:
  - 1000s of typefaces in use
  - but, do well only on commonly occurring ones
- Multi-lingual

  there exists *no single* technology that is readily retargetable to *any new* language
- Modest successes:

  e.g. fixed forms, telephone bills

Retargetable OCR Systems

- User assists the system:
  - provides models specific to the document
  - sacrifices full automation,

    but <u>gains accuracy & versatility</u>
- PARC research: we can model:

  language, typefaces, layout, image quality
- Large improvements: 2-10x fewer errors!
- But, are users willing to go to the trouble?

PARC's Document Image Decoding

- Explicit formal stochastic models of
  - **text generation**: language
  - **image rendering**: typefaces, page layout
  - **image quality**: 'salt-&-pepper' noise

  (combined in a single FS Markov network)
- Integrated search for optimal 'decoding'
  - MAP criterion
  - search: Viterbi and variants
- Algorithmic optimizations for speed *only*
- Extensible to grey-scale, other languages
- Trainable using sample page images w/ 'truth'

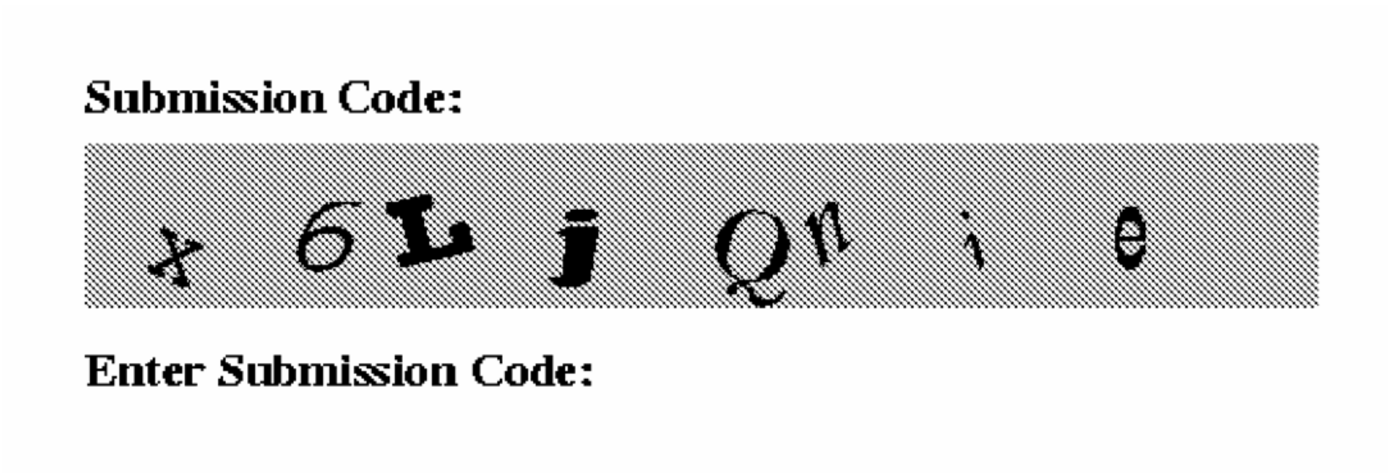
Legacy Document Conversion

- Large repositories
  - of long or similar printed documents
- Paper $\rightarrow$ ASCII, XML, Unicode, ...
  - scanning, recognition, manual correction
- Established service-bureau business
  - manual correction is expensive
- Research need recognized more and more:
  - NSF/DARPA/NASA Digital Library Initiative
  - ACM+IEEE Portal proposal
  - More in the near future….

UC Berkeley Digital Library Project

- Depts of CS & SIMS:
  – ‘Reinventing Scholarly Information Dissemination’
  – testbed: ‘CalFlora’ botanical website
  – users: Botanical scholarly community
- PARC is participating:
  – experimental BookScanner for rare & fragile books
  – whole book’s images up on the Web
  – next: a PDA field guide!

DIA Impact on Web Security: e.g. Altavista's AddURL filter



- 1997: noticed robotic abuse of 'Add-URL' feature
- 2000: Andrei Broder *et al* tried "ransom note" filter

  … reduced "spam add_URL" by "over 95%"

Alan Turing (1912-1954)

1936 a universal model of computation

1940s helped break Enigma (U-boat) cipher

1949 first serious uses of working computer

including plans to read printed text

(he thought it would be easy)

1950 proposed test for machine intelligence

“CAPTCHAs”:

Completely Automated Public Turing Tests to Tell Computers & Humans Apart

(M. Blum, L. A. von Ahn, J. Langford, et al, CMU SCS)

- challenges can be generated & graded automatically

  (i.e. the judge is a machine)
- accepts virtually all humans, quickly & easily
- rejects virtually all machines
- resists automatic attack for many years

  (even assuming that its algorithms are known?)

NOTE: the machine administers, but cannot pass the test!

PARC/UCB's PessimalPrint: exploiting image degradations

OCR machines fail when:

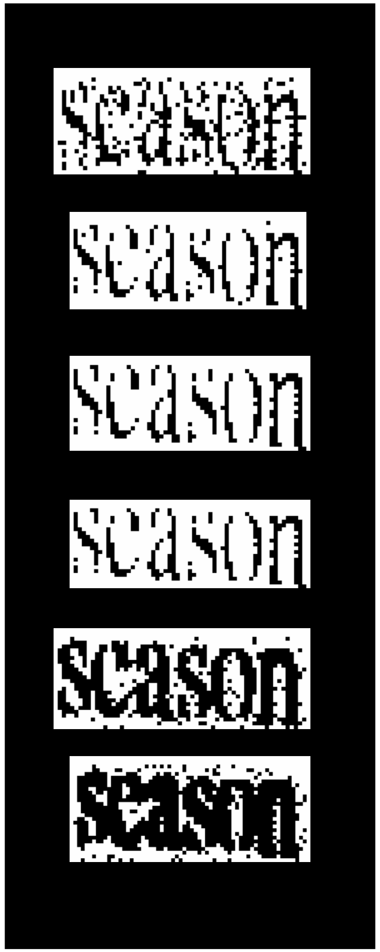
- **blur** = 0.0

  & **threshold** $\in$ 0.02 - 0.08

- **threshold** = 0.02

  & any value of **blur**

… but **people read them easily**



OCR outputs

- ~~~.I~~~
- ~~i1~~
- N/A
- N/A
- N/A
- ~~I~~

Lots of Open Research Questions

- **What are the most intractable obstacles to OCR?**

  segmentation, occlusion, degradations, …?

- **Under what conditions is human reading most robust?**

  linguistic & semantic context, Gestalt, style consistency…?

- **Where are ‘ability gaps’ located?**

  quantitatively, not just qualitatively

- **How can we generate challenges *within* the ability gaps?**

  fully automatically

  an indefinitely long sequence of distinct challenges

DIA Nagging Research Questions

- Can human performance be matched?
  - or exceeded?!
- Can engineering be fully automated?
  - e.g. by training: obviate $$ custom solutions
- Can systems be easily retargeted?
  - escape from tiny niche markets
- Can systems adapt autonomously?
  - avoid training, tuning