Reconocimiento de Escritura

Daniel Keysers

Image Understanding and Pattern Recognition
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany

Mar-2007

Outline

Outline

OCR

- OCR = Optical Character Recognition
- steady progress in OCR since the mid-fifties
- 1975: IBM Optical Page Reader cost over 3,000,000 US$ and displaced several dozen keypunch operators
- applications
  - home or office use
  - forms processing
  - address reading
  - conversion of large archives of text to computer-readable form

OCR

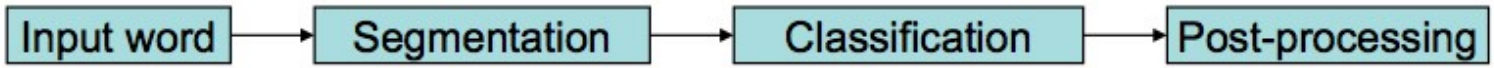
Although researchers have worked on the problem of OCR for at least thirty years, there has been a renewed interest in OCR technology in the recent years. This is partly due to

- the increasing need for efficient information storage and retrieval,
- the increasing need for cross-language information access, and
- the dramatic drop in scanner prices
- [large scale digitization projects]

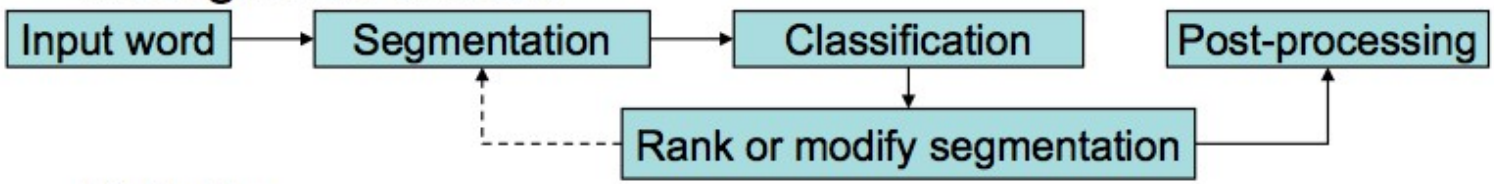
(Kanungo, 1998)

OCR Paradigms

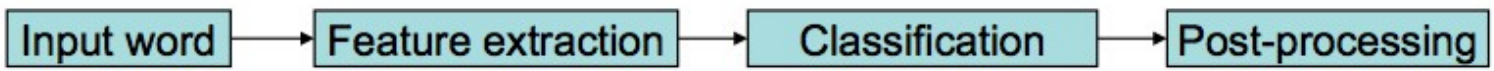
- OCR paradigms: [Casey 96]
  - Dissection (Segmentation driven OCR):



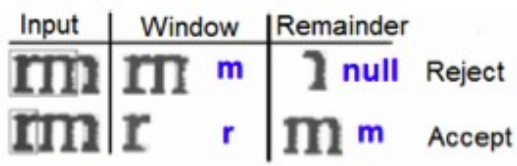
  - Recognition driven:



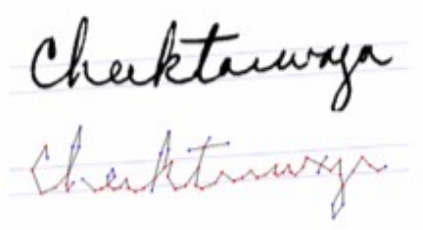
  - Holistic:



Segmentation driven



Recognition driven



Holistic

OCR

building blocks of an OCR system:

- noise removal
- skew estimation
- page segmentation & line detection (= layout analysis)
- segmentation
- character classifier
- language modeling

What are the main differences to a handwriting recognition system?

OCR

even 99% accuracy
= 30 errors on a typical printed page of 3000 characters

"in almost every application, either the OCR results must be corrected by a human operator or a significant fraction of the documents are rejected in favor of operator entry" (Nagy$^+$ 2000)

Comercial OCR

2 Browser and Design Testing

There are multiple implementations of HTML rendering engines; some common ones are Microsoft's Internet Explorer, Mozilla's Gecko, Apple's Safari, Opera's browser, and KDE's KHTML. Each of these render web pages differently due to bugs and incomplete specifications of web standards. Common defects are missing text, text that is unintentionally rendered overlapping, text that unintentionally overlaps graphical elements, bad font substitutions, bad spacing, and unreadable choices of foreground and background colors.

Our approach to this problem is to render the HTML into an image-based representation and then subject the image-based representation to OCR (including layout analysis). Common rendering problems can be detected by comparing the HTML input against the OCR output. For example, incorrect rendering due to missing text, overlapping text, bad font substitutions, and text in invisible colors can be flagged by detecting text that is present in the original HTML but missing in the OCR output. Incorrect layouts can be detected by comparing the paragraph structure and reading order of the original HTML against the layout analysis output.

We automate this process by using user interface scripting support. Our initial prototype is implemented on Apple Macintosh OS X, where we use a combination of AppleScript, the Firefox ScreenGrab extension, and the Safari Snagit extension to automatically capture web pages with different browsers, versions, and browser settings, and to send the captured page to be processed by the OCR system; analogous technologies exist for Windows and Linux. This way, a large collection of web pages can be rendered, analyzed, and verified without the need for operator intervention in different browsers and browser versions. The approach can detect HTML layout problems without prior assumptions about layout engines and for browsers returning a wide variety of layouts; the approach correctly distinguishes incorrect and correct layouts even in the presence of JavaScript and style sheets.

The same approach can be used for checking HTML layouts against design rules. Design rules for HTML are intended to ensure readability, accessibility, and easy interpretation by readers, as well as to ensure correct representations of organizational identity. Design rules specify such features as minimum font sizes, acceptable fonts and color choices, and minimal spacings between logical groupings of page elements.

3 Phishing and Search Engine Spam

Phishing (www.antiphishing.org) is a problem in which an adversary attempts to obtain personal and private information by creating E-mails and web pages that belong to a trusted organization (e.g., the recipient's bank), but actually transmit the information to the adversary. Closely related to phishing is search engine spam, where a web site will present HTML content to a search engine that gives indications to the search engine that the web site contains relevant information about a popular topic but actually renders as

2 Browser and Design Testing

There are multiple implementations of HTML rendering engines; some common ones are Microsoft's Internet Explorer, Mozilla's Gecko, Apple's Safari, Opera's browser, and KDE's KHTML. Each of these render web pages differently due to bugs and incomplete specifi- cations of web standards. Common defects are missing text, text that is unintentionally rendered overlapping, text that unintentionally overlaps graphical elements, bad font sub- stitutions, bad spacing, and unreadable choices of foreground and background colors.

Our approach to this problem is to render the HTML into an image-based representa- tion and then subject the image-based representation to OCR (including layout analysis)...

Commercial OCR

44 LECTURE 4. EUROPEAN OPTIONS IN COMPLETE MARKETS

Indeed, it follows from (3.5′) in view of (4.39) that

$$\begin{aligned}(4.40)\qquad \mathcal{E}_n^{-1}(U)\,X_n^{\pi_n} &= (1-\alpha)\mathbf{C} + \sum_{k=1}^{n}\mathcal{E}_k^{-1}(U)\,\gamma_k^{\pi}\,S_{k-1}\,(\rho_k - r_k)\\ &= (1-\alpha)\mathbf{C} + \sum_{k=1}^{n}\mathcal{E}_k^{-1}(U)\,S_{k-1}\,\gamma_k^{\pi}\,(\rho_k - r_k)\\ &\quad - \sum_{k=1}^{n}\mathcal{E}_k^{-1}(U)\,S_{k-1}\,\varphi_k\,\mathbf{C}\,(\rho_k - r_k)\\ &= (1-\alpha)\mathbf{C} - (1-\alpha)\,\mathbf{C} + M_n^{\varphi} - \mathbf{C}\,\mathbf{M}_n^{\varphi}.\end{aligned}$$

From (4.40),

$$\mathcal{E}_N^{-1}(U)\,X_N^{\pi_n} = \mathcal{E}_N^{-1}(U)\,f - \mathbf{C}\,\mathbf{I}_{\{Z_N<\lambda\}}$$

and hence

$$\mathcal{E}_N^{-1}(U)\,X_N^{\pi_n} \geq \mathcal{E}_N^{-1}(U)\,f - \mathbf{C}.$$

The last inequality means that $\pi_n \in \mathrm{SF}(f,N)$.

Further, it follows from (4.38) that

$$\begin{aligned}(4.41)\qquad \mathbf{P}^*\{X_N^{\pi_n} \geq f\} &= \mathbf{P}^*\{f - \mathbf{C}\,\mathcal{E}_N(U)\,\mathbf{I}_{\{Z_N<\lambda\}} \geq f\}\\ &= \mathbf{P}^*\{\mathbf{I}_{\{Z_N<\lambda\}} \leq 0\} = \mathbf{P}^*\{Z_N \geq \lambda\} = (1-\alpha).\end{aligned}$$

Finally, we get from (4.38) and (4.41) that

$$\begin{aligned}(4.42)\qquad \mathbf{P}\{X_N^{\pi_n} \geq f\} &= \mathbf{E}^*\,\mathbf{I}_{\{X_N^{\pi_n}\geq f\}}Z_N\\ &\geq \mathbf{E}^*\,\mathbf{I}_{\{X_N^{\pi_n}\geq f\}}\mathbf{I}_{\{Z_N\geq\lambda\}}Z_N\\ &\geq \lambda(1-\alpha) \geq 1-\alpha.\end{aligned}$$

The relations (4.41) and (4.42) show that the condition (4.35) holds for the strategy π_n, and hence π_n is an α-$((1-\alpha)\mathbf{C}, f, N)$-hedge.

What has been obtained shows that it is possible to hedge a contingent claim *with a specified probability* $(1-\alpha)$. Further, *the initial funds can be reduced by the amount* $\alpha\,\mathbf{C}$, though with a *risk* α the accepted contingent claim cannot be repaid.

PROBLEMS

4.1. Prove that on a no-arbitrage (B,S)-market we have for a standard European option to buy (sell) that $\mathbf{C}(N_2) \geq \mathbf{C}(N_1)$ (respectively, $\mathbf{P}(N_2) \geq \mathbf{P}(N_1)$) when $N_2 \geq N_1$.

4.2. Prove that the fair price $\mathbf{C} = \mathbf{C}(N, S_0, K)$ of a standard European option to buy, where N is the exercise time, S_0 is the initial price of a share, and K is the exercise price, has the following properties:

a) $\mathbf{C}(S_0, K)$ is monotone in S_0 and K;

d) $\mathbf{C}(S_0, K)$ is convex in S_0 and K;

c) $\mathbf{C}(\lambda S_0, \lambda K) = \lambda\,\mathbf{C}(S_0, K)$ for $\lambda > 0$.

Indeed, it follows from (3.5') in view of (4.39) that (4.40) S k=I fc=i = (1 - a)C - (1 - a) C + From (4.40), ﷺ and hence The last inequality means that TTQ 6 SF(f,N). Further, it follows from (4.38) that (4.41) P'{X# > /} = P*{/ -C﷿N(U) I{ZN f} = P'{I{ZN< 0} = P*{ZN > A} = (1 - a). Finally, we get from (4.38) and (4.41) that (4.42) P{X^>/}=E'IW﷿ > A(l-a) > 1-a. The relations (4.41) and (4.42) show that the condition (4.35) holds for the strategy na, and hence TTQ is an a-((l - a)C, /, A^)-hedge. What has been obtained shows that it is possible to hedge a contingent claim with a specified probability (1 ﷿ a). Further, the initial funds can be reduced by the amount a C, though with a risk a the accepted contingent claim cannot be repaid. PROBLEMS 4.1. Prove that on a no-arbitrage (B, 5)-market we have for a standard Euro- pean option to buy (sell) that C(7V2) > C(Ni) (respectively, P(JV2) > P(M)) when 4.2. Prove that the fair price C = C(N, So, K) of a standard European option to buy, where N is the exercise time, So is the Initial price of a share, and K is the exercise price, has the following properties: a) C(S0, K) is monotone in So and K; d) C(So, K) is convex in So and K; c) C(\S0,XK) = A C(S0,K) for A > 0.

Commercial OCR

Liber Scholasticus;

OR,

AN ACCOUNT OF THE

FELLOWSHIPS, SCHOLARSHIPS,

and

EXHIBITIONS,

at the

Universities of Oxford and Cambridge;

BY WHOM FOUNDED,

AND WHETHER OPEN TO NATIVES OF ENGLAND AND WALES,
OR RESTRICTED TO PARTICULAR PLACES AND PERSONS:

ALSO, OF SUCH

Colleges, Public Schools, Endowed Grammar Schools,

CHARTERED COMPANIES OF THE CITY OF LONDON,

CORPORATE BODIES, TRUSTEES, &c.

AS HAVE UNIVERSITY ADVANTAGES ATTACHED TO THEM,
OR IN THEIR PATRONAGE.

WITH APPROPRIATE INDEXES AND REFERENCES.

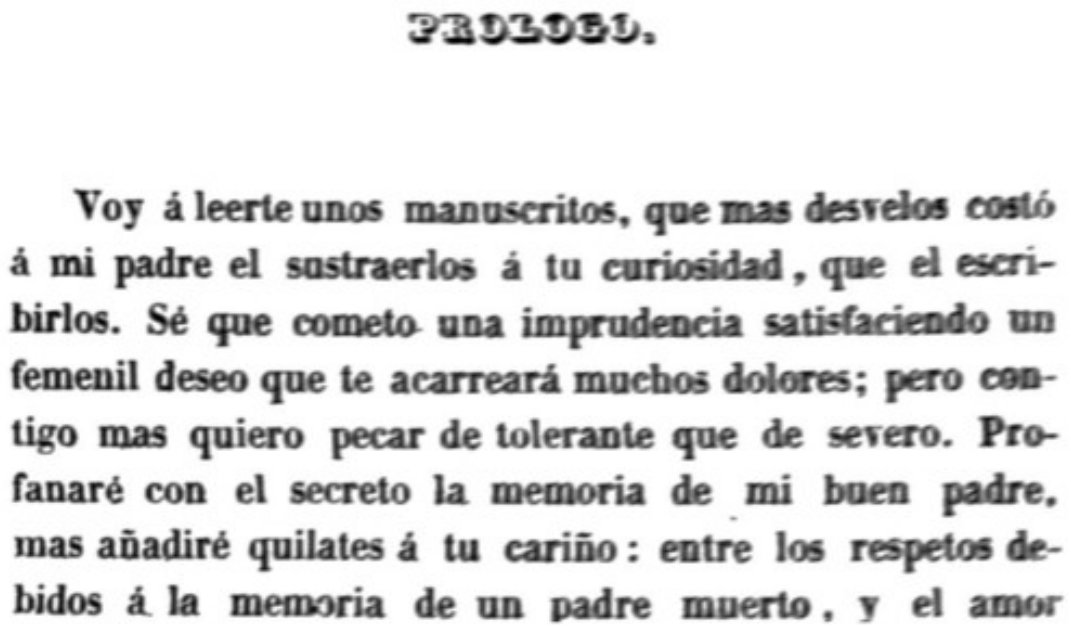
LONDON:
PRINTED FOR C. J. O. & F. RIVINGTON,
ST. PAUL'S CHURCH-YARD, AND WATERLOO-PLACE, PALL-MALL.

MDCCCXXIX.

OR,
AN ACCOUNT OP THE
FELLOWSHIPS, SCHOLARSHIPS,
and
EXHIBITIONS,
at the
attttonvitto of <C2><A9>Tforfc anft
<E2><82><AC>amfitiU0
BY WHOM FOUNDED,
J.VJ> UHKrilKK OPEff TO IfATIfES OF
SNOLAND AND WALES,
Ott RKiTRICTEU TO PARTICULAR
PLACES AND PERSONS;
ALSO, OF SUCH
CoKrgrs, IJutlir $rf)ool6, Kniutotti (Grammar
5rf)ool
CHABTERED COMPANIES OF THE CITY
OF LONDON,
CORPORATE BODIES, TRUSTEES, &c.
At BArS OXIrESSJTY ADrANTAOES
ATTACHED TO TBEX,
OS IN THEM PATRONAGE.
WITH APPROPRIATE INDEXES AND
REFERENCES.
^LONDON:
PRINTED FOR C. J. O. & F. RIV1NGTON,
. PAl L's Clif RCII.YAlin, AMl
WATERLOO.PLACE, PALUMALL.

MDCCCXXIX.

Commercial OCR



PRÓLOGO.

Voy á leerte unos manuscritos, que mas desvelos costó á mi padre el sustraerlos á tu curiosidad, que el escribirlos. Sé que cometo una imprudencia satisfaciendo un femenil deseo que te acarreará muchos dolores; pero contigo mas quiero pecar de tolerante que de severo. Profanaré con el secreto la memoria de mi buen padre, mas añadiré quilates á tu cariño: entre los respetos debidos á la memoria de un padre muerto, y el amor

â¬*MInv-
Toy aleertennof n^m^ritn. qaeva* desTdos eosto
4 mi padre d snstnerlos a tu oniosidad, qae d eseri-
birlos. Se" qne cometa ana impradtncia iilirfirirÂ«dn on
femenil deseo qne te aearreara modiM dokns; pcro ew-
tigo mas quiero pecar de tolerant* qne de wrcro. Pra-
fanart COD el secrete la memoria de mi boen padre.
mas anadirt qoilates a tu carioo: eatre 1Â« respeto* de-
bidos a. la memoria de on padre nmerlo, j d amor

Systems

commercial:

- Abbyy (FineReader)
- Nuance/Scansoft (OmniPage)
- Océ (RecoStar)
- Iris (ReadIris)
- many smaller vendors

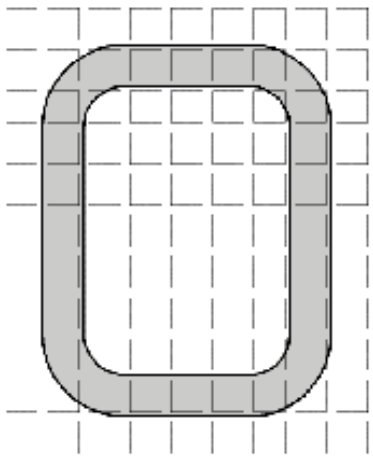
open source:

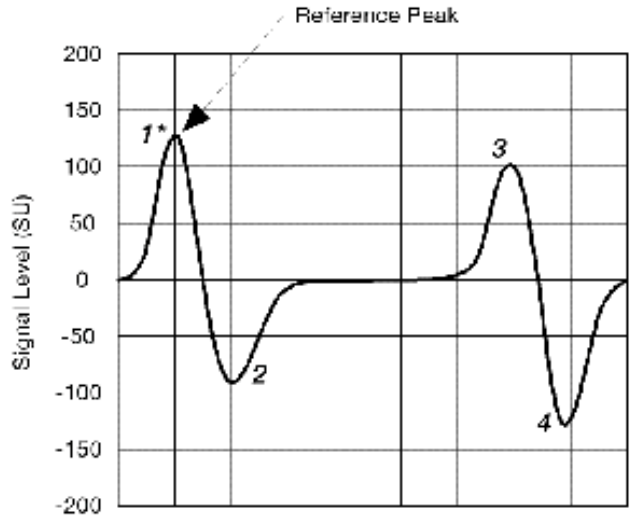
- ocrad
- gocr
- Tesseract
- OCRopus

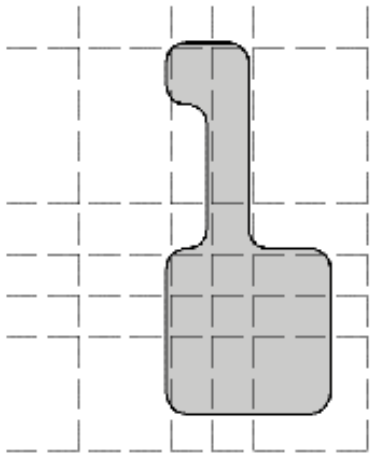
Outline

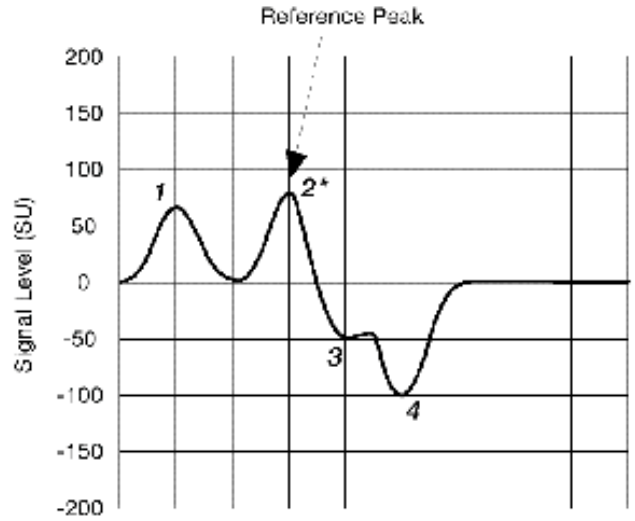
OCR fonts

If you can, try to make the problem easier. This is usually easier than improving the classifier.









OCR fonts

two groups of OCR fonts

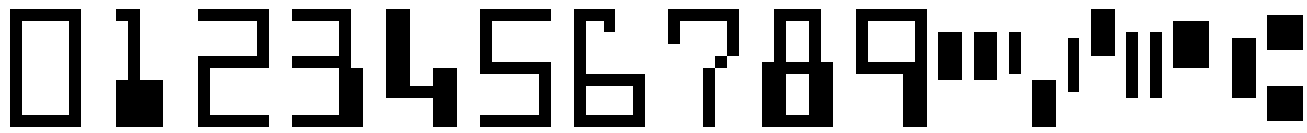
- Magnetic Ink Character Recognition (MICR)
- Optical Character Recognition (OCR)
- artificial distinction

history:

- financial world
- bar-code not easily readable by humans
- magnetic ink preferred

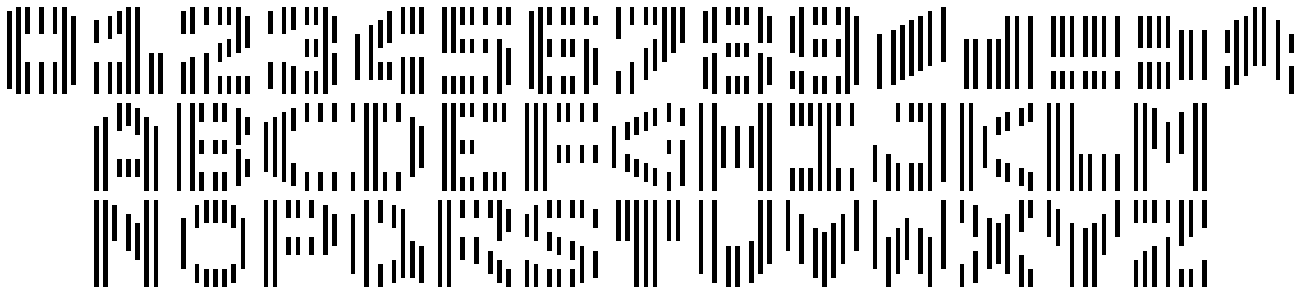
(source: D. Winter)

OCR fonts - E13-B



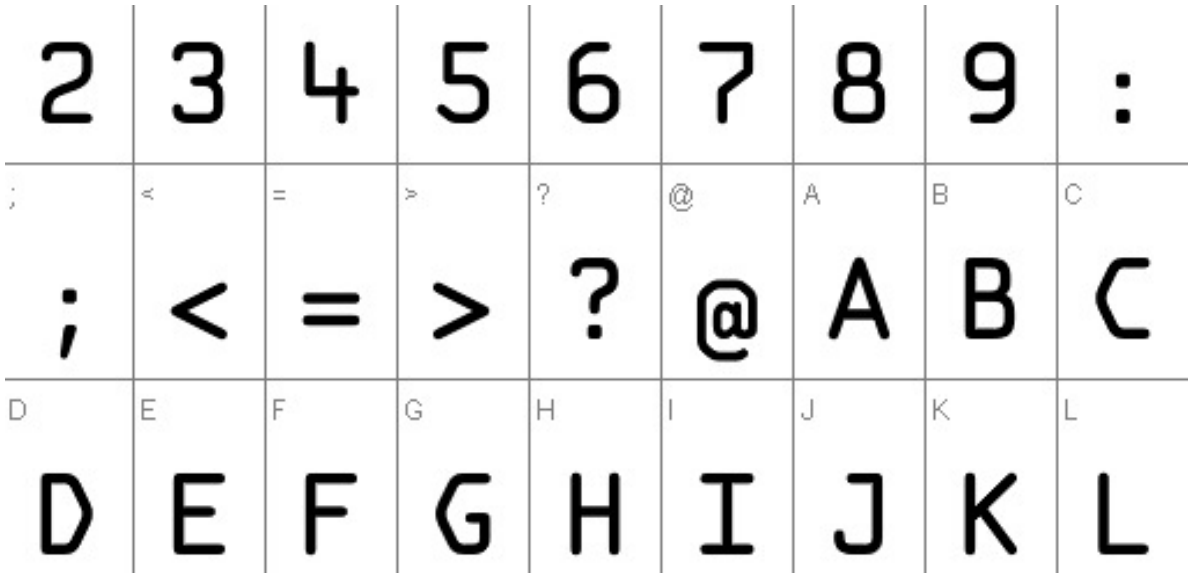
- first font used in automated banking
- digits and four special symbols used in banking" "dash", "amount", "on*us" and "transit"
- The font was specially designed so that magnetic pulses would be read unambiguously. That is the reason for some of the heavy black features of some symbols. For this purpose a grid of 7 by 11 squares was used that either had to be white or black. The filling was done so that a magnetic scan would give a pulse signal that was very distinctive.

OCR fonts CMC-7



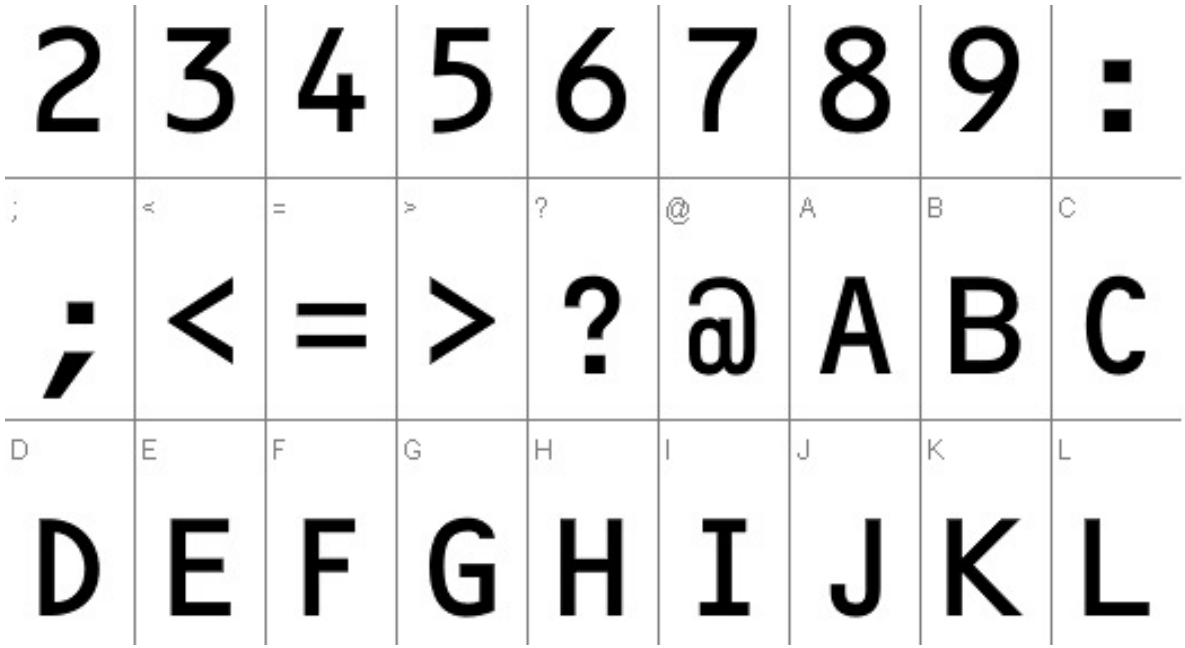
- digits, the letters, and five special symbols
- can also be seen as a barcode:
  each symbol encoded by seven vertical bars separated by small or large spaces

OCR fonts - OCR-A



- end of the sixties: full character recognition
- first font: OCR-A

OCR fonts - OCR-B



- accompanies EAN barcodes
- more symbols exist

Output Representation: hOCR Format

- proposed open standard for representing OCR results
- motivation: existing formats have limitations in
  - multi lingual capabilities
  - typographic phenomena
  - separate formats for intermediate and final results
- goal: reuse as much existing technology as possible
- main idea: represent various aspects of OCR output
  - logical structuring
  - typesetting
  - character information
  - etc.
- HTML microformat

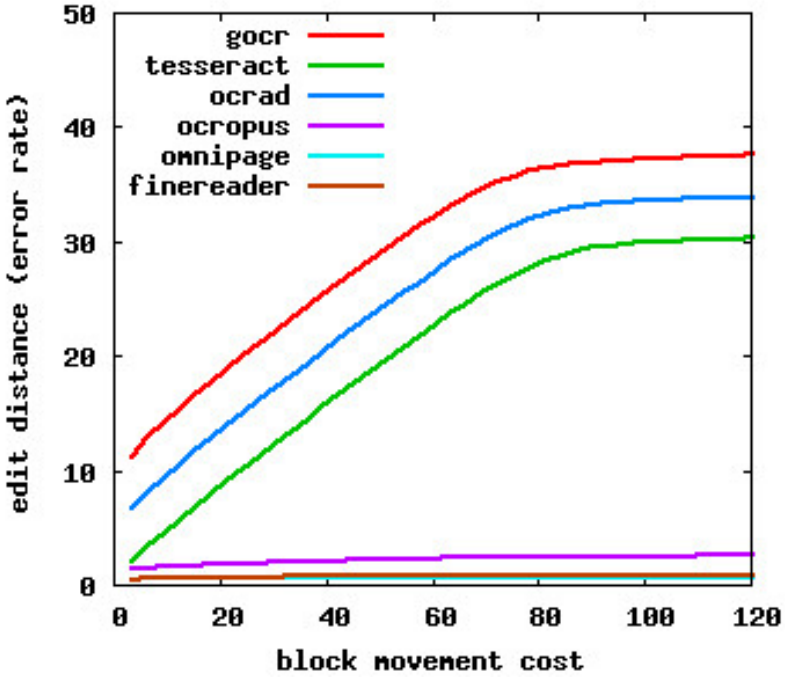
Assessment of OCR Error Rate

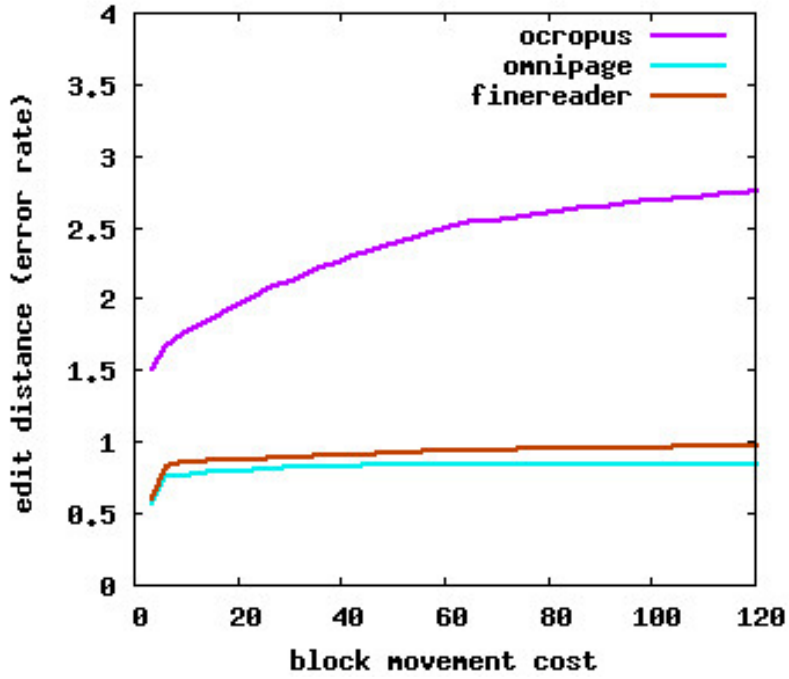
- usually: edit- or Levenshtein distance (cp. Speech Recognition)
- in the presence of reading order errors: allow block movements at certain cost (cp. Machine Translation)
- interesting problem: predict OCR accuracy from a set of document image measurements

OCRopus Open Source OCR System

- Layout Analysis → see previous lecture
- Character Recognition Engine:
  - based on Tesseract → discussed here
  - based on segmentation → see lecture on handwriting recognition
- flexible architecture

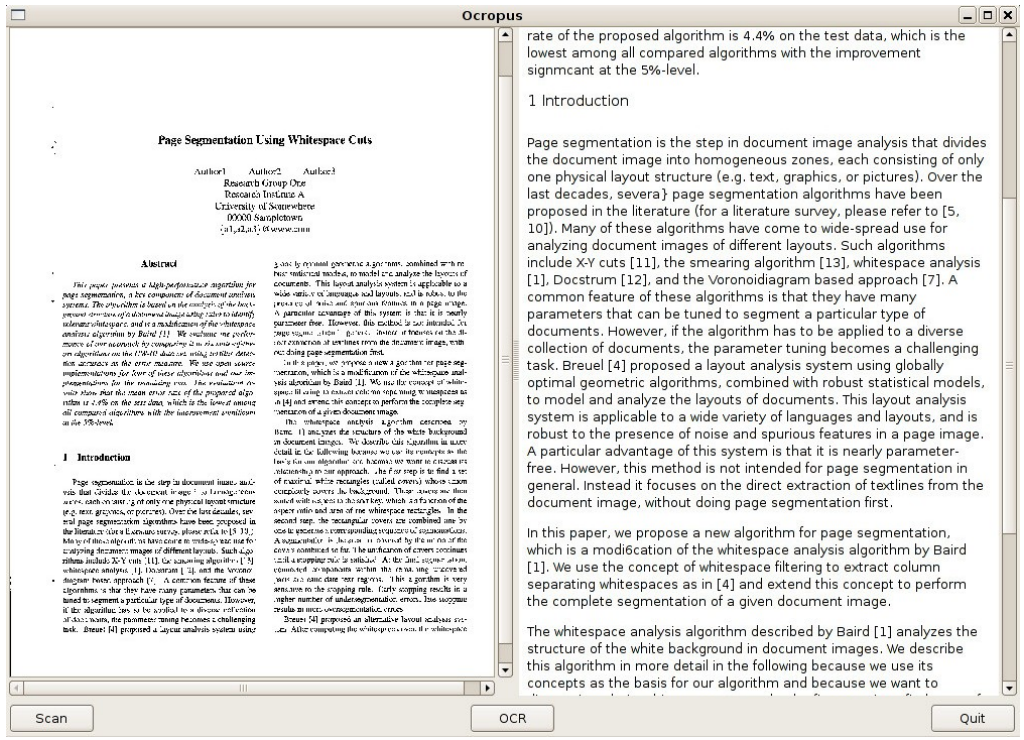
OCRopus Open Source OCR System





- 'OCRopus' is the DFKI open source OCR system
- framework for layout analysis and OCR integration
- DFKI layout analysis + DFKI or HP-labs 'Tesseract' OCR
- preliminary evaluation on 18 documents, 15K words
- goals: improving adaptivity to new fonts and layouts

OCRopus Open Source OCR System — Release



- first version (technology preview) to be released soon
- other application: screen OCR

screen OCR challenges

antialiasing

binarization works poody

very small x-height

Screen OCR

- motivation:
  - image-based HTML analysis
  - image-based cut and paste
- character recognizers:
  - HMMs
  - Tesseract

Outline

History

R. Smith: 'An overview of the Tesseract OCR Engine', submitted to ICDAR 2007, personal communication.

- developed at HP between 1984 and 1994
- obtained good results at the 1995 UNLV Annual Test of OCR Accuracy
- then, development was stopped while other commercial OCR engines improved
- In late 2005, HP released Tesseract for open source.

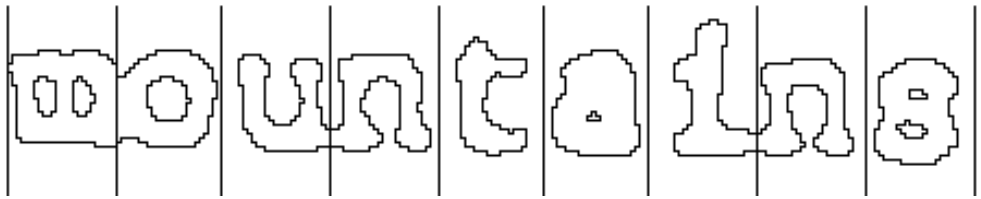
Architecture

- Layout Analysis was separate, therefore not included
- (OCRopus closes this gap by including the possibility to use DFKI layout analysis with the Tesseract recognition engine)
- only supports US-ASCII

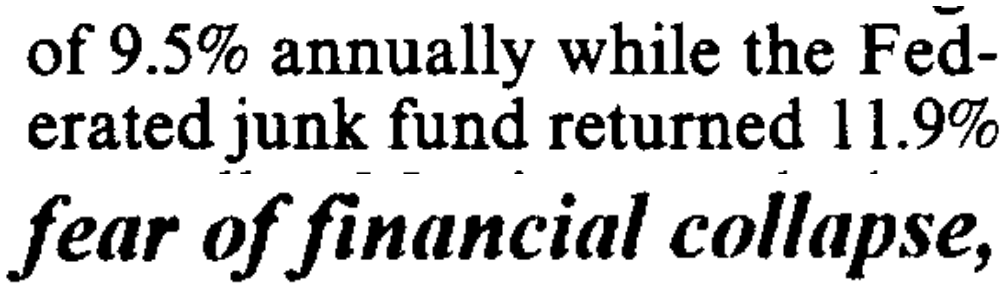
- connected component analysis
- stored as outlines ('blobs')
  $\longrightarrow$ enable recognition of inverse text easily
- blobs $\longrightarrow$ text-lines

Chopping into Words

distinguish fixed pitch and proportional text



fixed-pitch: chopping simple



proportional: Measure gaps in a limited vertical range between the baseline and mean line. Spaces close to the threshold are made fuzzy, so that a final decision can be made after word recognition.

Adaptation Using Two Passes

first pass:

- attempt to recognize each word
- 'good' words are used as adaptation data

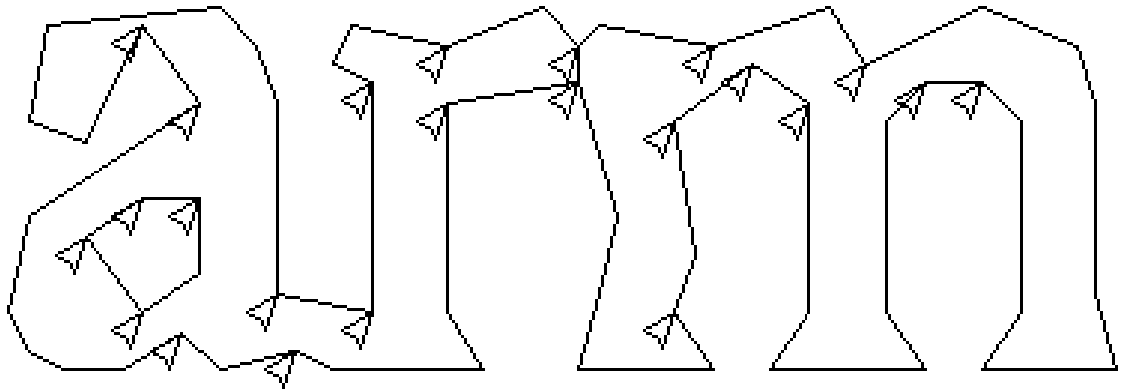
second pass:

- recognize words that were not recognized well again
- use adaptation data

Line Finding

- does not need de-skewing
- filter large and small blobs out
- fit remaining blobs to parallel text line model
- re-assign left out blobs
- fit baselines as splines

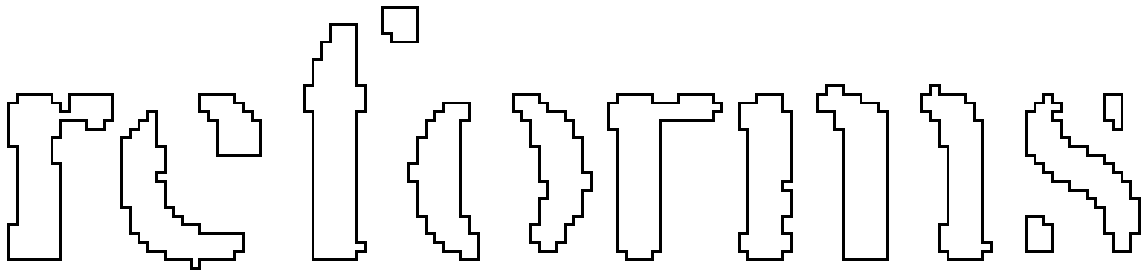
Segmentation of Touching Characters



While the result from a word is unsatisfactory, Tesseract attempts to improve the result by chopping the blob with worst confidence.

Candidate chop points are found from concave vertices of a polygonal approximation of the outline, and may have either another concave vertex opposite, or a line.

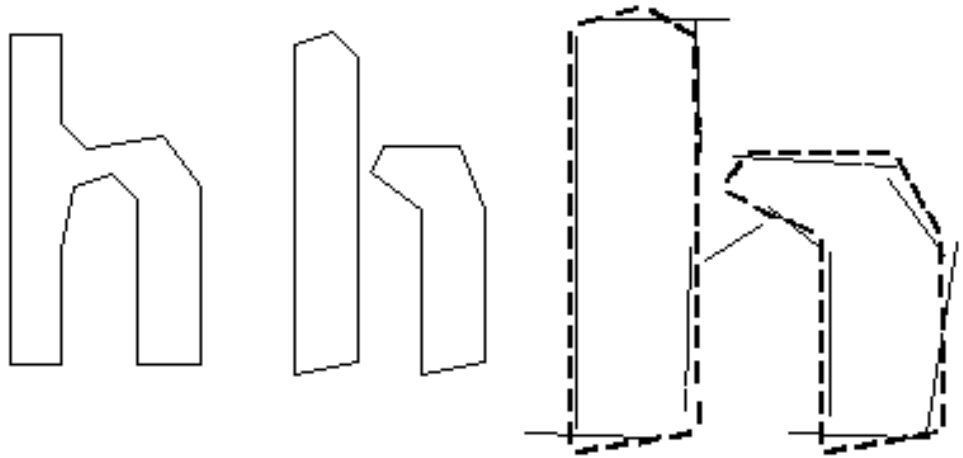
Combine Broken Characters



A* search of the segmentation graph of possible combinations of the maximally chopped blobs into candidate characters without actually building the segmentation graph, but instead maintaining a hash table of visited states

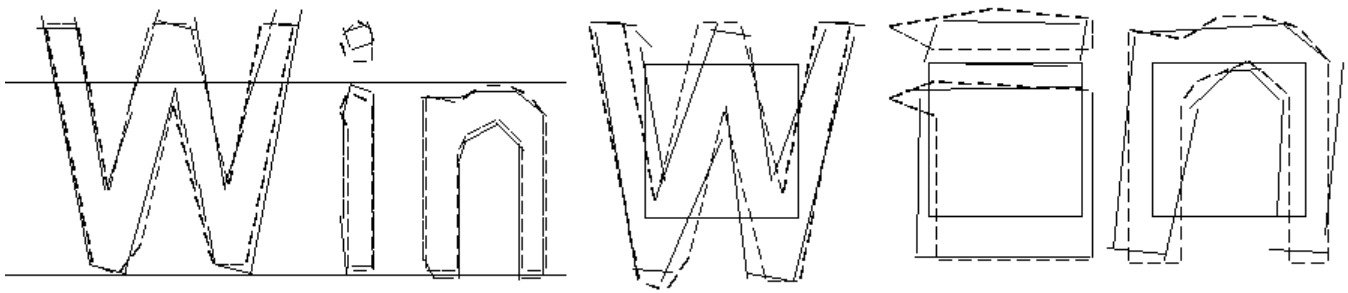
character classifier can recognize broken characters directly

Character Classifier



- match outlines of test many-to-one
- test features 3-dimensional, (x, y position, angle), typically 50-100
- prototype features are 4-dimensional (x, y, position, angle, length), typically 10-20 features in a prototype
- hierarchical classification and use of lookup-tables for speed-up

Normalization



baseline and moment normalization used in adaptive and static classifier

Training Data

- no need for broken characters in training
- 20 samples of 94 characters from 8 fonts in a single size, but with 4 attributes (normal, bold, italic, bold italic), making a total of 60,160 training samples
- other classifiers: often more than 1,000,000 training samples

Linguistic Analysis

very basic linguistic analysis:
look-up in different dictionaries

Outline

Major Classes of OCR Error Sources

Imaging Defects	Similar Symbols	Punctuation	Typography
Heavy Print	Similar Vertical Symbols	Commas and Periods	Italics and Spacing
Light Print	Other Similar Symbols	Quotation Marks	Underlining
Heavy and Light Print		Special Symbols	Shaded Backgrounds
Stray Marks			Reverse Video
Curved Baselines			Unusual Typefaces
			Very Large Print
			Very Small Print

(Nagy$^+$ 2000)

Examples

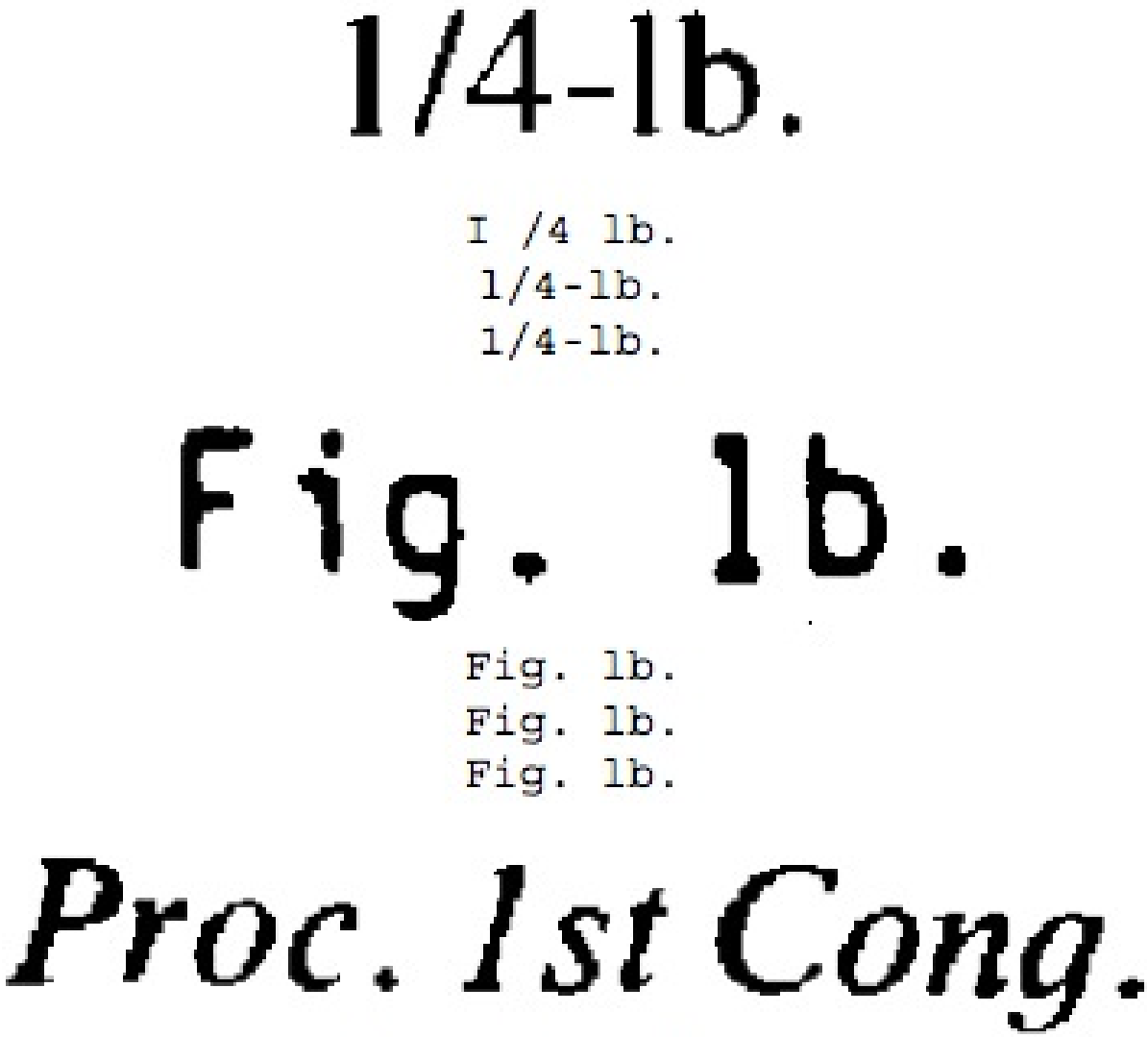
RecDmENSATIONS
RECO-HENATIDNS
REC-ENDATIOffS

Fig. 1a. The Siamese **M**'s baffle all three systems.
Small pieces are missing from the second **M**, the first **N**, and the **D**.

site.
sites
sitea

Fig. 1b. A few pixels can make all the difference. Each system recognizes the first **s**,
but only one can identify the second. Perhaps comparing the two **s**'s would help.

Examples


1/4-lb.
I /4 lb.
1/4-lb.
1/4-lb.
Fig. 1b.
Fig. 1b.
Fig. 1b.
Fig. 1b.
Proc. 1st Cong.
Proc. I st Cong.
Proc. I st Cong.
Proc. 1st Cong.

Examples

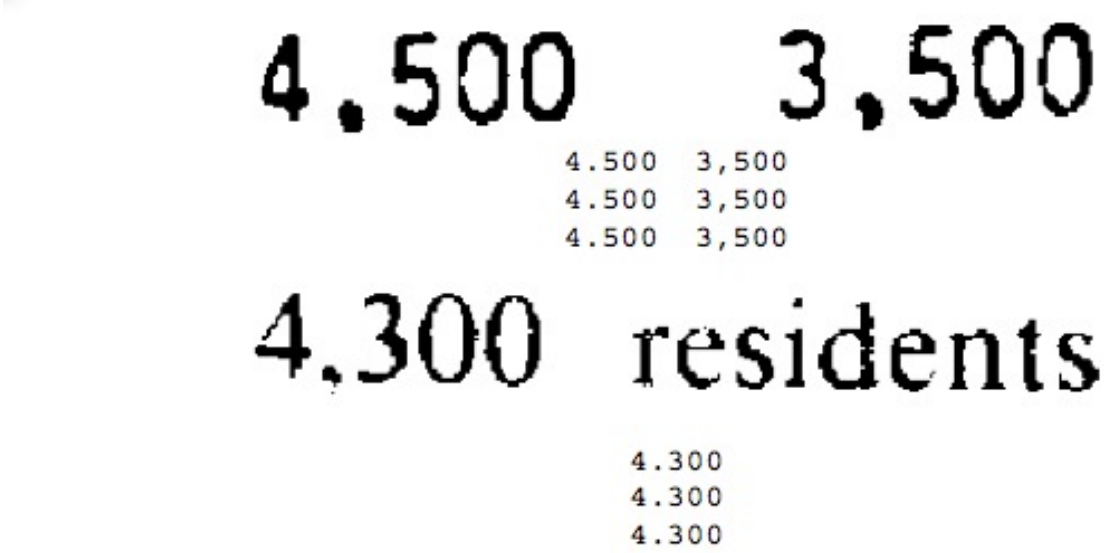


Fig 3. All three devices conspire for a thousand-fold reduction of **4,500** and **4,300**, but a few extra pixels save **3,500**.

Examples

Fig. 4. Kerned italics cause a variety of problems, include the omission of interword blanks. Underscoring is also difficult. All the devices are bewildered by the enormous flourish on the capital **R** (truly a 'majuscule'), and by the z-like appearance of the lower-case **r**.

Potential Sources of Improvement

(Nagy$^+$, 2000)

- improved image processing
- adaptation of the classifier to the current document
- multi-character recognition
- increased use of context
- [combination of multiple OCR results]

Adaptation

- OCR systems generally optimized for average performance over large sets of characters of different
  - fonts
  - sizes
  - scan qualities
- improve performance of character recognizer by using style information

Adaptation Approaches

- modeling the sample distribution [Breuel 2001]
  - each document contains only a single style
  - modeling sample distribution as a mixture of Gaussians
- hierarchical Bayesian approach [Mathis et al. 2002]
  - each document contains a small number of fonts
  - estimation of prior distributions of a style variable
- style constrained classifiers [Sarkar et al. 2005]
  - style consistency constraint is hidden variable
  - combination of class and style represented by Gaussian mixture model
  - each mixture component trained by estimating it directly from set of samples from a specific class and style
- maximum likelihood linear regression [Senior et al. 1997]
  - based on work on speaker adaptation [Legetter 1995]
  - adaptation using linear transformations