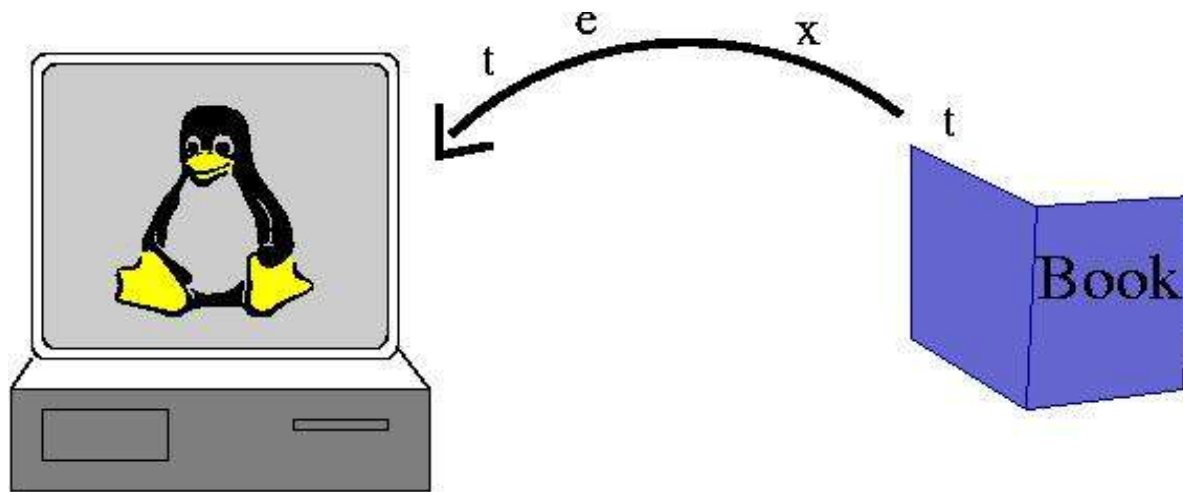


GOCR

Optical Character Recognition



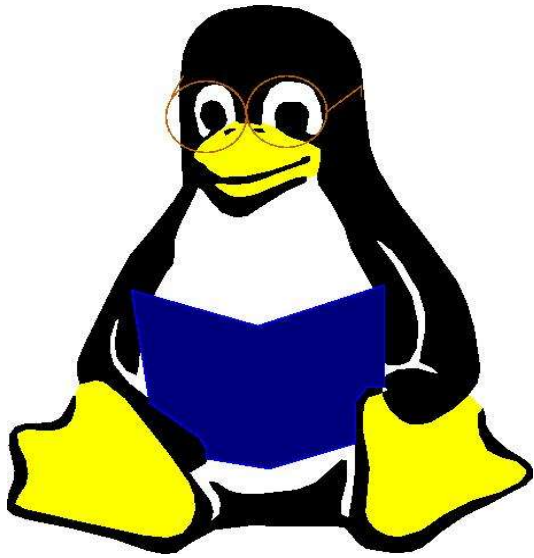
GOOCR, what can it do for you?

- **converting images** with printed text **to text** format
- making printed text accessible for text utils
- **sources:** scans, screenshots, photos?, pdf's, edocs
- **destination:** ASCII/UTF-text, HTML/XML, TeX

GOOCR, the goal ...

just **take** it and **use** it

GOOCR

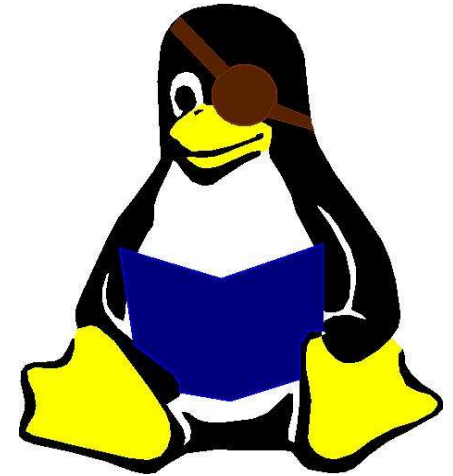


Features?

- portability, flexibility (OS, c, pnm, text only)
- font independent, no training, executable only
- autodetection (can be switched off)
- barcodes

GOOCR

Restrictions (v0.40)?

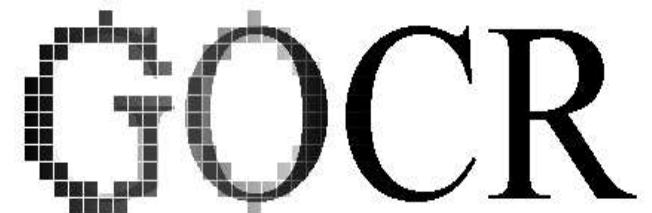


- only clean black on white (high res, no dust etc.)
- nonrotated, simple layouts, (printed) text only
- only latin font with some extensions (de, fr, se)
- only 1D barcodes
- slow, ...

GOOCR

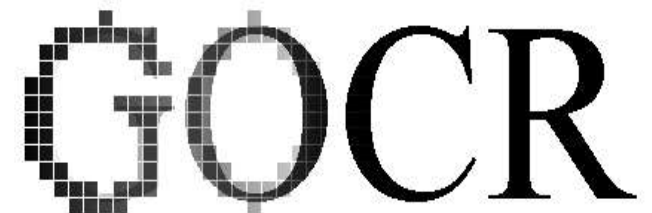
Alternatives?

- require mouse and X
 - Clara OCR, OCRchie
- experimental
 - OCRE (rough FFT)
 - locr (fast, simple database)

The logo for GOOCR, where the 'G' is rendered in a pixelated, blocky font, and the 'OOCR' is in a standard serif font.

How to measure quality?

- Quality of recognition? (speed, resources?)
- Recognition rate in percent? 99%..99.999% ?
- Depends strongly from the pattern. Standards?
- Don't care?
- Popularity? (press, downloads, feedback)

The logo for GOOCR, where the 'G' is rendered in a pixelated, blocky font, and the 'OOCR' is in a standard serif font.

Greatest success?

- First implementation of RFC 1149, April 2001
 - IP over avian carrier (Bergen Linux User Group)
 - carrier pigeon internet protocol (CPIP)



<http://de.wikipedia.org/wiki/Bild:Postduif.jpg>

GOOCR

Joerg Schulenburg, LinuxTag 2005 - GOOCR

2nd greatest success?

- asprise.com's OCR Java SDK 2.1, April 2005
 - selling it for \$998..\$2,998 (but GPL violation)
 - advertises his program as very accurate without being afraid of malcontented buyers
- Web-excerpt:

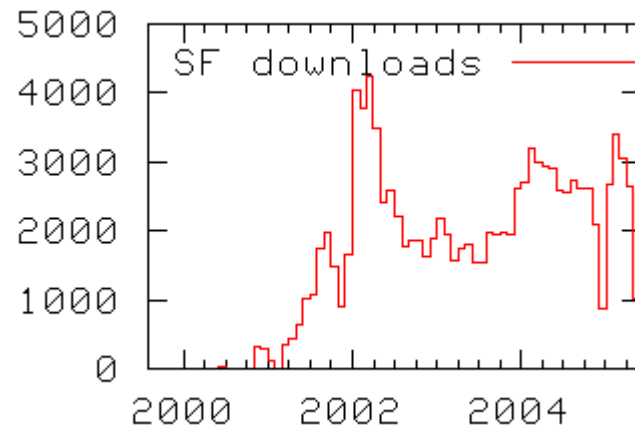
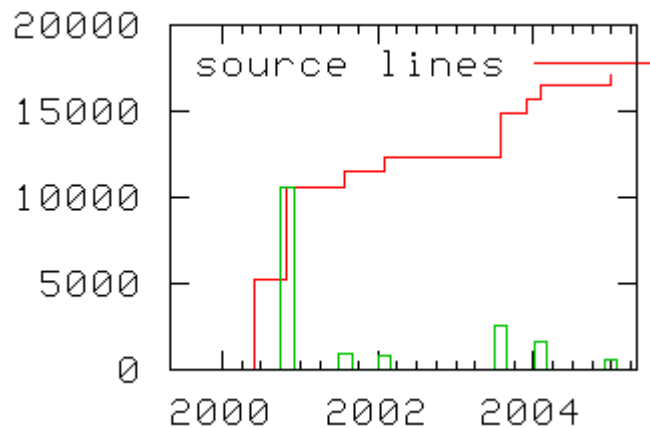
■ Features of Asprise OCR SDK:

1. **Highest Level of Accuracy**

Asprise OCR can easily recognize difficult documents of poor image quality;

2. **Excellent Format Retention**

Code history?



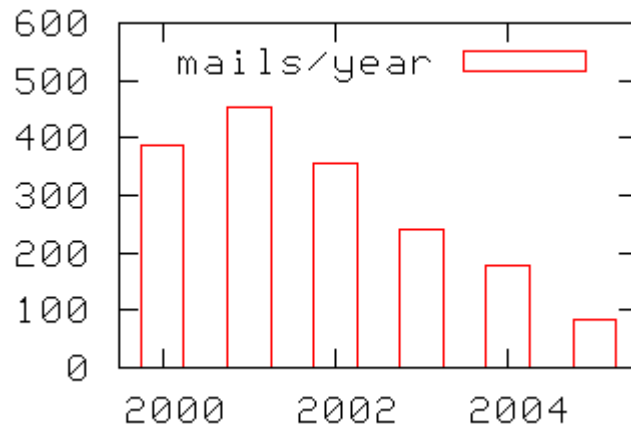
- v0.1 autumn 1998
- v0.2.1 Mar2000 @freshmeat
May2000 1000 downloads
- v0.2.5 Jun2000 @SF
2001 (SuSE-7.1)
- v0.3.5 Feb2002 (SuSE-8.1)

18000 code lines (v0.40)

When can we release v1.0?

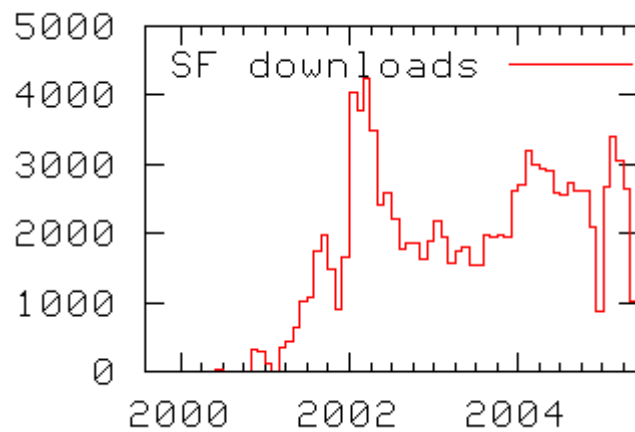
GOOCR

Email traffic?



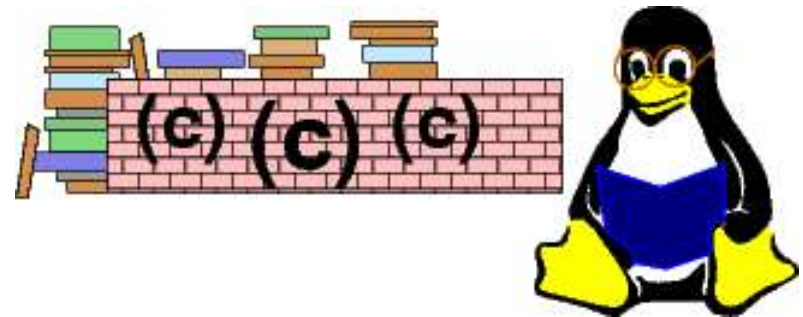
- every 2nd day a ocr message
- 20x more Spam

Final state is no emails but downloads?



GOOCR

Problems?

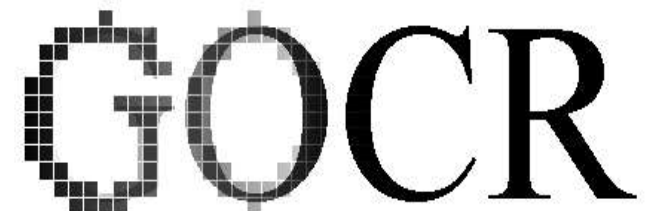


- Copyright and sample sources (big problem)
- not enough time for coding

GOOCR

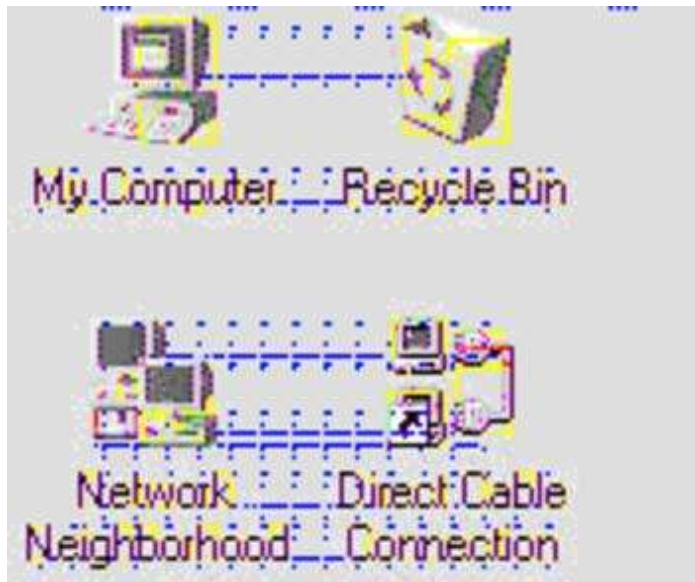
How does it work?

- preprocessing
 - threshold value detection
 - box detection, zoning, line detection
 - sorting and melting, dust, pictures, ...
- call ocr engine (3 engines, 2 experimental)
 - repeated for unknown chars
- postprocessing (XML, TeX, UTF, ASCII)

The logo for GOOCR features the word "GOOCR" in a large, black, serif font. The letter "G" is rendered in a pixelated, blocky style, while the letters "O", "O", "C", and "R" are in a standard serif font.

How does it work?

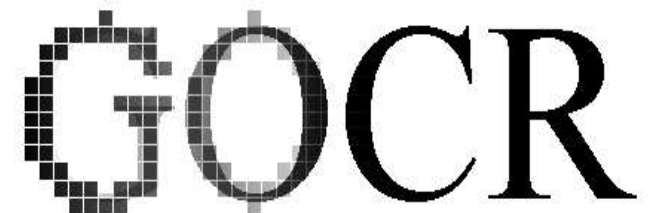
- preprocessing
 - threshold value detection
 - box detection, zoning, line detection
 - sorting and melting, dust, pictures, ...



GOOCR

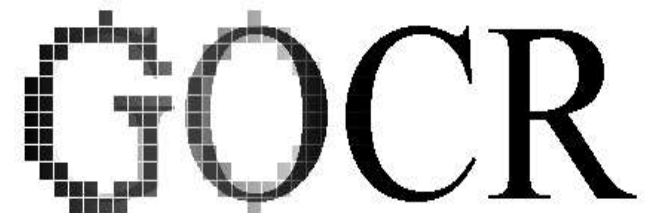
Why is it not working?

- no idea ... P-)
- very **simple algorithms** written by an OCR amateur, making the program usable at a early state of coding
- most code written around midnight?
- its **alpha** code and will be better (**v0.xx**)

The logo for GOOCR, where the 'G' is rendered in a pixelated, blocky font, and the 'OOCR' is in a standard serif font.

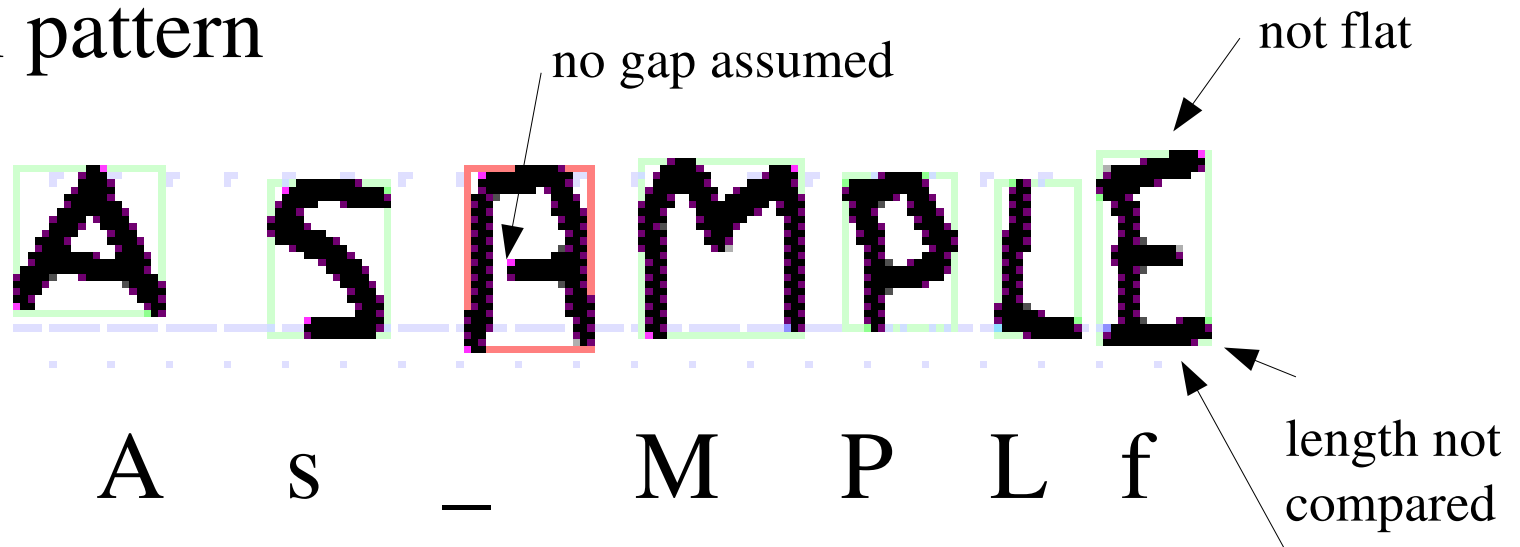
Looking inside ...

- assume no **colors**, black on white only
- assume no rotation, same *font*, all characters are `s e p a r a t e d`
- try to repair if assumptions are hurt (can fail)
- every char is recognized empirically based on its pixel pattern
- lot of possibilities to improve GOCR

The logo for GOCR, where the 'G' is rendered in a pixelated, blocky font, and the 'OOCR' is in a standard serif font.

Looking inside ...

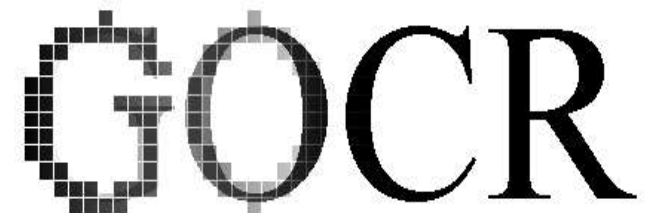
- every char is recognized empirically based on its pixel pattern



GOOCR

Useful for accessibility?

- OCR is important for visible impaired people
- Linux is cheap and flexible
- but good and flexible OCR is missing
- GOCR will try to serve that needs (feedback?)

The logo for GOCR, where the letter 'G' is rendered in a pixelated, blocky font, while the letters 'O', 'C', and 'R' are in a standard serif font.

Challenges?

- www.seeingwithsound.com/ocr.htm
- screenshots (small fonts and graphic mixed)
- signposts on photos (webcam shots)



Challenges?



Photo: A. Karwath, commons.wikimedia.org

- . . .
- car plates?

GOOCR

next steps - colors

outer text Helvetica 20

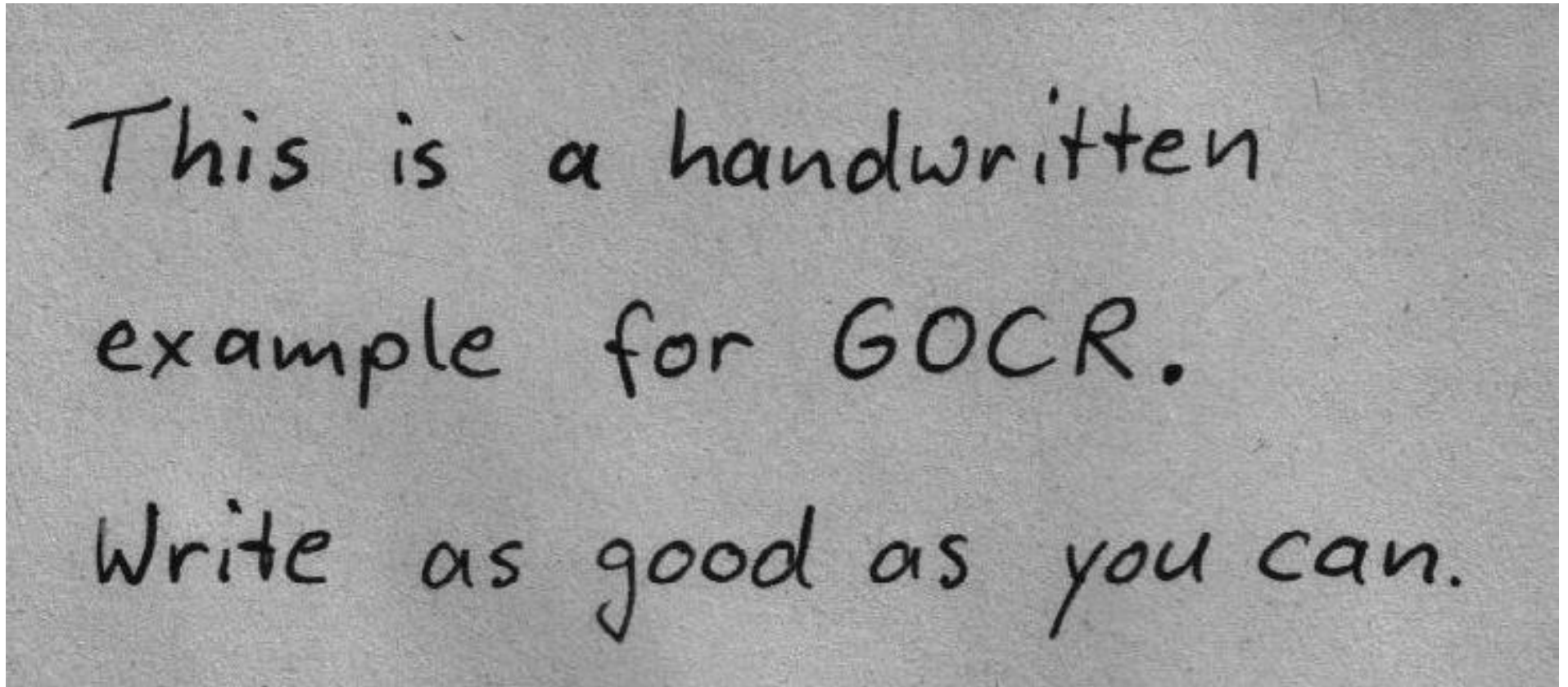
Inner-1 Times-Roman 20

This is a box in a box.

step 2

step 3 box

next steps – hand written

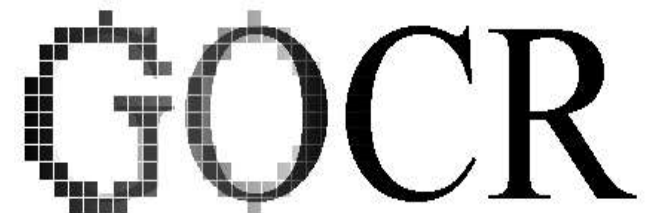


next steps – rotated text

Times–Roman 12
The text is rotated by
45 degree.

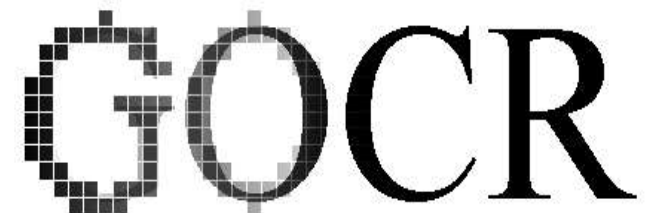
Timeline?

- improve recognition (stepwise, on demand)
- multi color for screenshots -- Aug05
- extract text from photos -- Oct05
- 2D barcodes (usefull?) -- Dez05
- vectorization and rotation – Mai06
- suggestions?

The logo for GOOCR, where the 'G' is rendered in a pixelated, blocky font, and the 'OOCR' is in a standard serif font.

How can you help?

- create copyright-free samples for typical problems (easy)
- write/find OCR related free available publications
- write patches (difficult)
- write a new+better OCR engine (most difficult)

The logo for GOOCR, where the 'G' is rendered in a pixelated, blocky font, and the 'OOCR' is in a standard serif font.