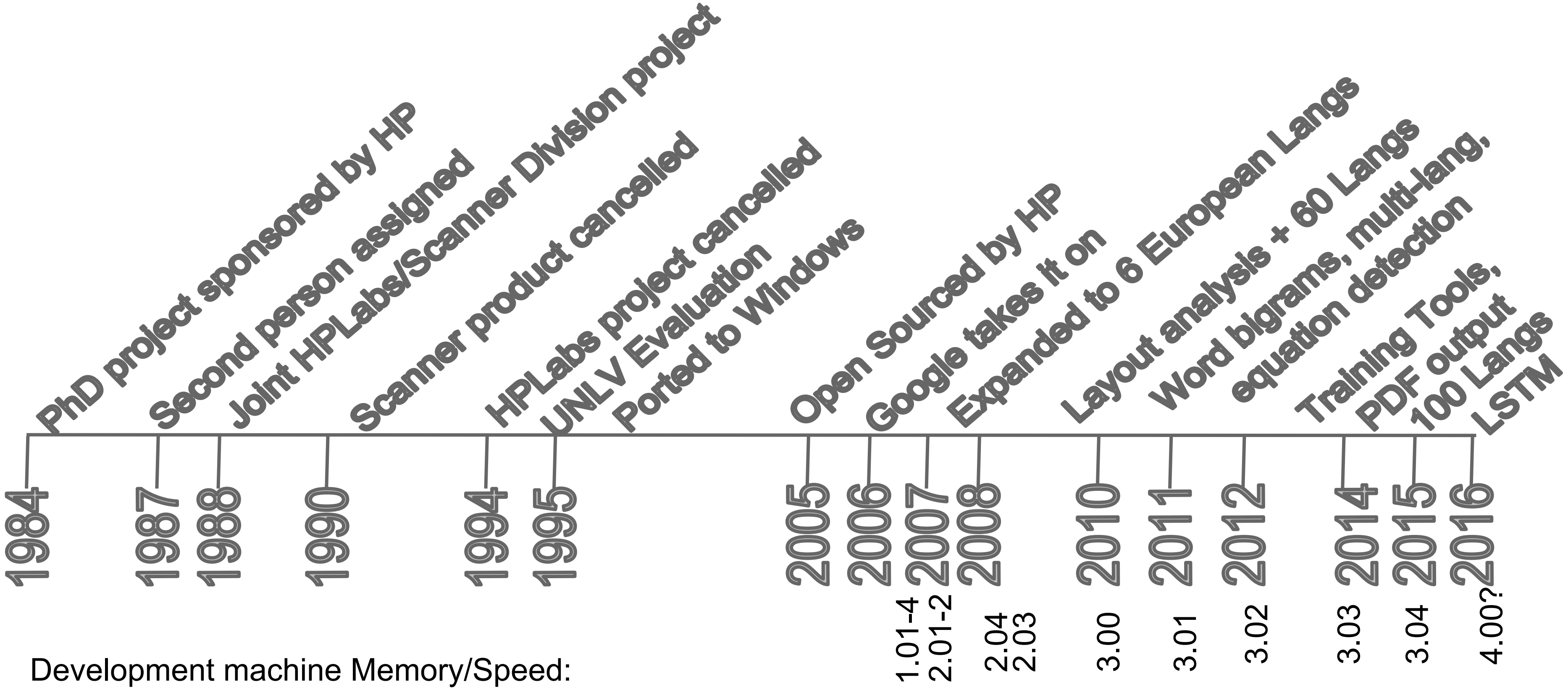
1. Motivation and History of the Tesseract OCR Engine

Lessons Learned from What Worked and What Didn't

Ray Smith, Google Inc.

What is Tesseract?

Tesseract Timeline



PhD project sponsored by HP
Second person assigned
Joint HPLabs/Scanner Division project
Scanner product cancelled
HPLabs project cancelled
UNLV Evaluation
Ported to Windows
Open Sourced by HP
Google takes it on
Expanded to 6 European Langs
Layout analysis + 60 Langs
Word bigrams, multi-lang,
equation detection
Training Tools,
PDF output
100 Langs
LSTM

1984
1987
1988
1990
1994
1995
2005
2006
2007
2008
2010
2011
2012
2014
2015
2016

1.01-4
2.01-2
2.04
2.03
3.00
3.01
3.02
3.03
3.04
4.00?

Development machine Memory/Speed:

4MB	16MB	64MB	256MB	3GB	24GB
10MHz	25MHz	100MHz	166MHz	3.5GHz	12 x 3.5GHz

https://github.com/tesseract-ocr/tesseract/blob/master/ChangeLog

The Stealth Era: 1984-1994

https://commons.wikimedia.org/wiki/File:F-117_Nighthawk_Front.jpg Public Domain Image

Office equipment trade-show:
Earl's Court, London late 1984

Dest has a scanner with built-in OCR with >99.5% accuracy.

By Guruleninn (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons
https://commons.wikimedia.org/wiki/File%3ADADF_(Canon_IR6000).JPG

Office equipment trade-show:
Earl's Court, London late 1984

Dest has a scanner with built-in OCR with >99.5% accuracy.

By Guruleninn (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons
https://commons.wikimedia.org/wiki/File%3ADADF_(Canon_IR6000).JPG

https://en.wikipedia.org/wiki/IBM_Selectric_typewriter#/media/File:IBM_Selectric_typeball.jpg
Public domain Image

Caveat: It only recognizes 8 fonts from IBM Selectric "golf-ball" typewriters.

Office equipment trade-show: Earl's Court, London late 1984

For documents conventional OCR can't handle, Kurzweil offers a better approach: ICR.

For almost two decades, optical character recognition systems have been widely used to provide automated text entry into computerised systems. Yet in all this time, conventional OCR systems have never overcome their in- ability to read more than a handful of type fonts and page formats. Propor- tionally spaced type (which includes vir- tually all typeset copy), laser printer fonts, and even many non-proportional typewriter fonts, have remained beyond the reach of these systems. And as a result, conventional OCR has never achieved more than a marginal impact on the total number of documents needing conversion into digital form.

http://www.iowablindhistory.org/blindhistory/tools-and-technology-optical-character-recognition-scanners Public domain image

Early Lessons:

Early Lessons:

- Be able to OCR your own sales literature!

Early Lessons:

- Be able to OCR your own sales literature!
- OCR should distinguish text from non-text.

Early Lessons:

- Be able to OCR your own sales literature!
- OCR should distinguish text from non-text.
- OCR should read white-on gray and black-on-gray as easily as black-on-white.

Early Lessons:

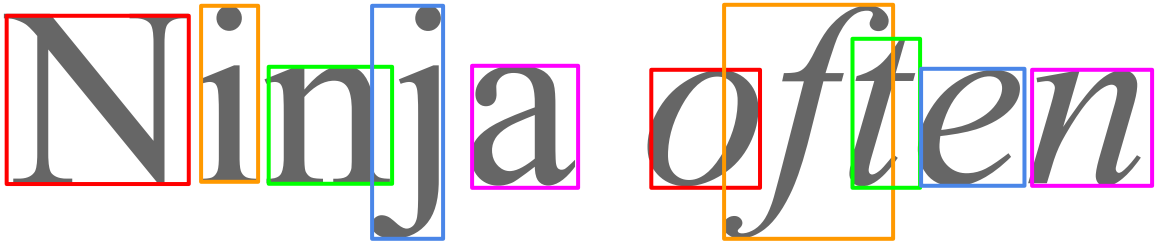
- Be able to OCR your own sales literature!
- OCR should distinguish text from non-text.
- OCR should read white-on gray and black-on-gray as easily as black-on-white.

Decision:

Extract outlines from grayscale images and features from the outlines.

- Profound impact.
- Key differentiator.

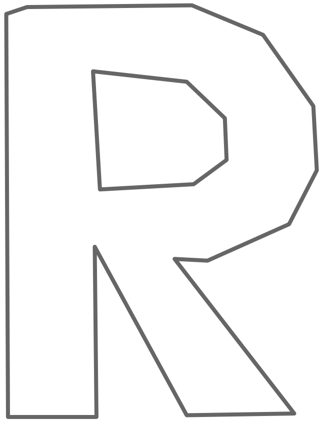
Connected component outlines vs bounding boxes:



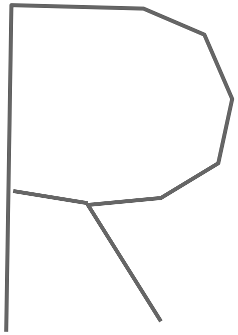
- Bounding boxes of characters often overlap without the characters actually touching.

How to extract features from Outlines?

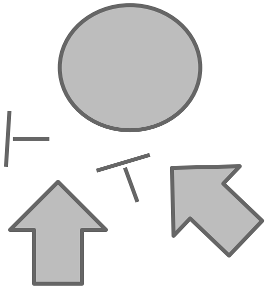
Outline



Skeleton

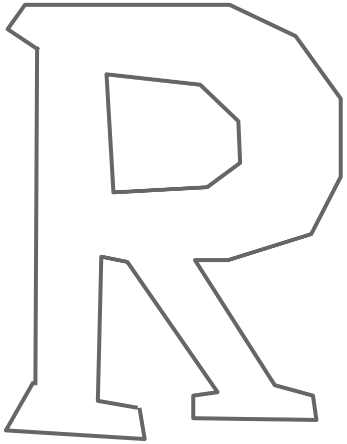


Topological Features

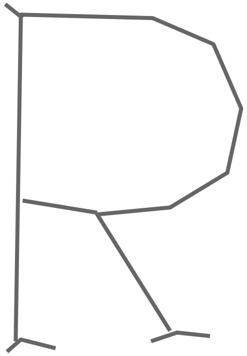


Skeletonization is Unreliable

Outline: Serifed



Skeleton: Decorated



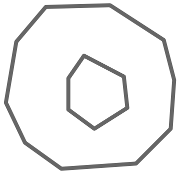
Arrrrh!

Lesson: If there are a lot of papers on a topic, there is most likely no good solution, at least not yet, so try to use something else.

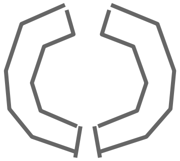
Topological features are Brittle

Damage to 'o' produces vastly different feature sets:

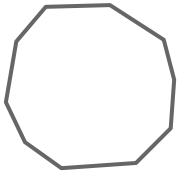
Standard 'o'



Broken 'o'



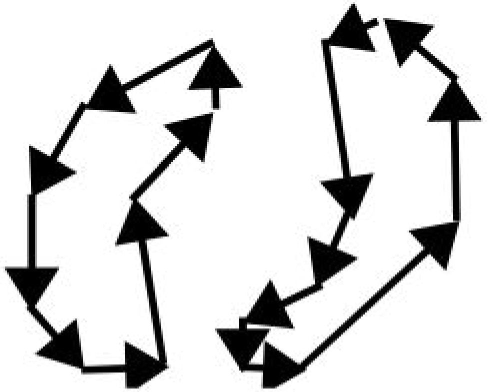
Filled 'o'



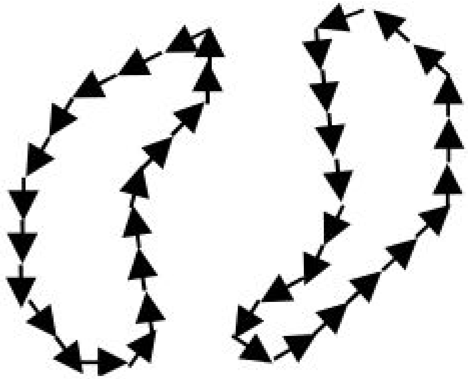
Lesson: Features must be as invariant as possible to as many as possible of the expected degradations.

Shrinking features and inappropriate statistics

Segments of the polygonal Approximation



Even smaller features



Statistical:

$$argmax(k)\prod_{l,i}\frac{1}{\sigma_{ijk}}\exp\left[-\frac{1}{2}\left(\frac{x_{il}-\mu_{ijk}}{\sigma_{ijk}}\right)^2\right]$$

Geometric:

$$argmin(k)\frac{1}{M+J_k}\left(\sum_{l,i}(x_{il}-\mu_{ijk})^2+\sum_{j,i}(x_{il}-\mu_{ijk})^2\right)$$

Lesson: Statistical Independence is difficult to dodge.

Now it's Too Slow!

~2000 characters per page x
~100 character classes (English) x
32 fonts x
~20 prototype features x
~100 unknown features x
3 feature dimensions

= 38bn distance calculations per page...

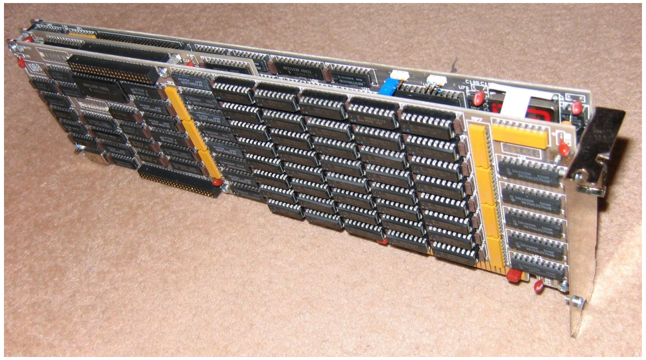
Now it's Too Slow!

~2000 characters per page x
~100 character classes (English) x
32 fonts x
~20 prototype features x
~100 unknown features x
3 feature dimensions

= 38bn distance calculations per page...
... **on a 25MHz machine.**

What Were Other OCR Systems doing in 1988-1990?

Calera Wordscan: Required a $2000 "real computer" add-on board to run a fixed dimension feature space on a PC (that used a 12.5MHz 80286 in those days):



By John Elliott (E-mail correspondence) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/)], via Wikimedia Commons https://commons.wikimedia.org/wiki/File%3APGC.jpg

Recognita: Decision tree using topological-like features. Very fast in software. Accuracy degraded very fast on poor quality.

Neural Networks: Were just becoming popular. Would there be a startup threat?

Testing

Summary:

- Small test sets are meaningless -> Get lots of compute power and data.
- Test on very different data to the training data.
- Test every change.
- What you measure improves -> Measure lots of dimensions.
- If it can break it will.
- Make your code faster while waiting for tests to run.
- Think/test out of the box...

Creative testing

Christmas 1992: Company shutdown for 10 days. What should we do with all that compute power that would otherwise be wasted?

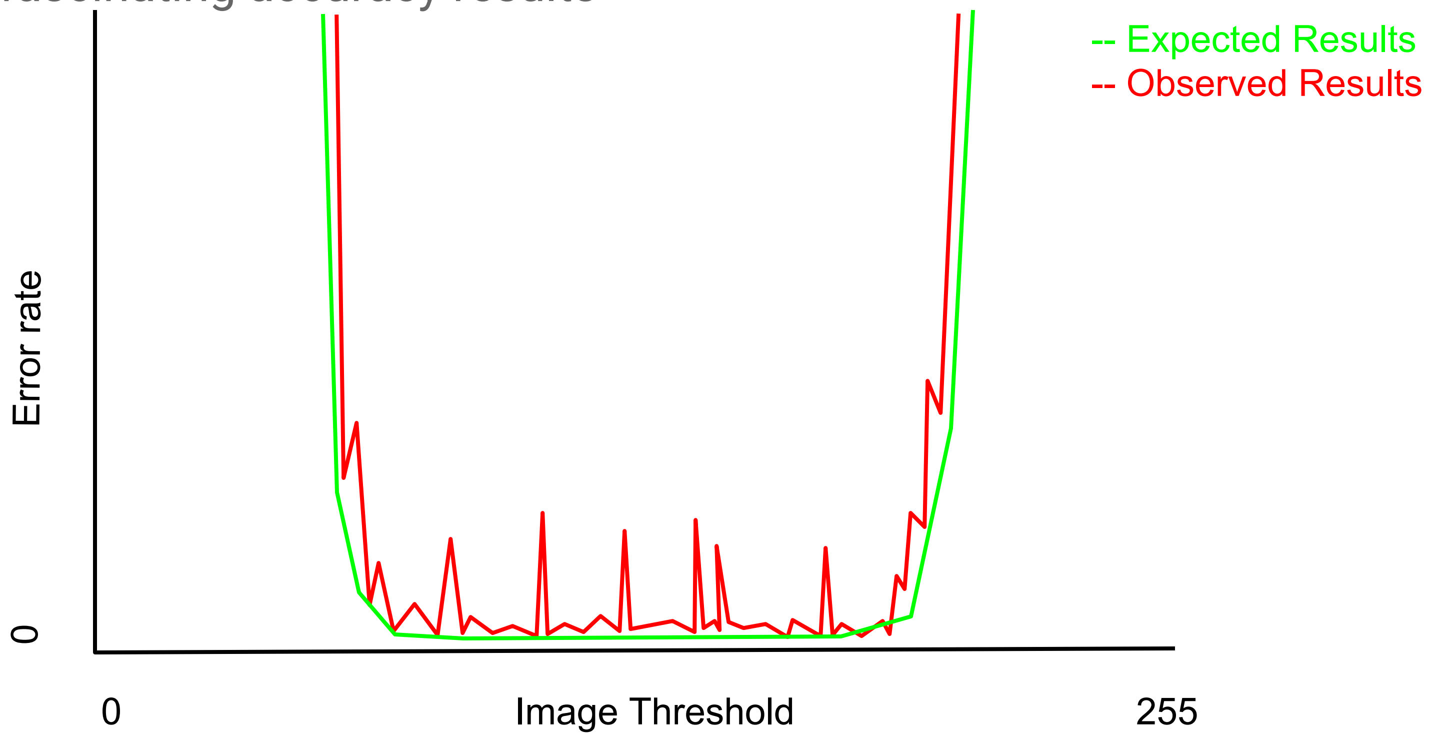
Creative testing

Christmas 1992: Company shutdown for 10 days. What should we do with all that compute power that would otherwise be wasted?

- Tune the hand-tuned parameters with a genetic algorithm.
- Process the 400 image test set thresholded at each possible threshold in the range 32-224

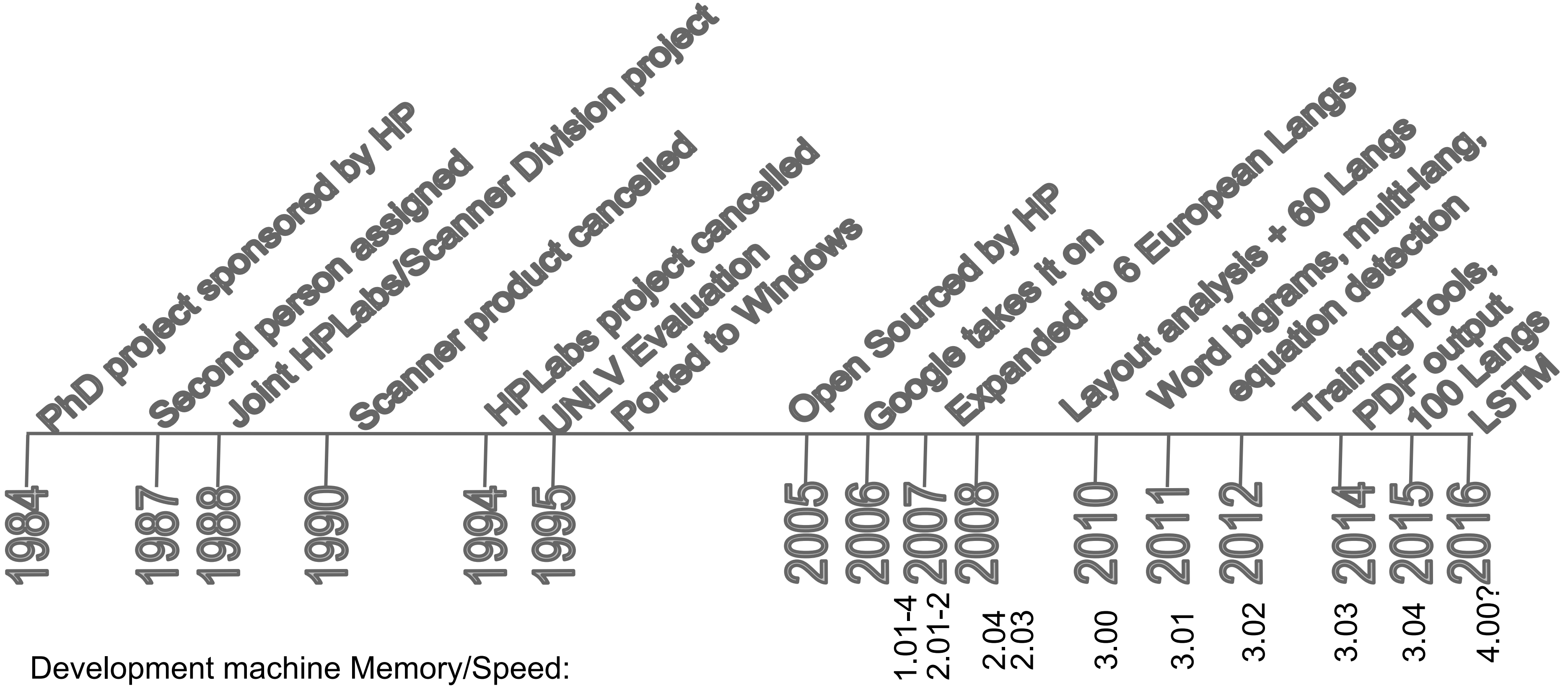
Thresholding Experiment Results

- Exposed several bugs
- Some fascinating accuracy results



Adaptive Thresholding for OCR: A Significant Test Ray Smith, Chris Newton, Phil Cheatle Personal Systems Laboratory
HP Laboratories Bristol HPL-93-22

Tesseract Timeline



PhD project sponsored by HP
Second person assigned
Joint HPLabs/Scanner Division project
Scanner product cancelled
HPLabs project cancelled
UNLV Evaluation
Ported to Windows
Open Sourced by HP
Google takes it on
Expanded to 6 European Langs
Layout analysis + 60 Langs
Word bigrams, multi-lang, equation detection
Training Tools,
PDF output
100 Langs
LSTM

1984 1987 1988 1990 1994 1995 2005 2006 2007 2008 2010 2011 2012 2014 2015 2016

1.01-4
2.01-2
2.04
2.03
3.00
3.01
3.02
3.03
3.04
4.00?

Development machine Memory/Speed:

4MB	16MB	64MB	256MB	3GB	24GB
10MHz	25MHz	100MHz	166MHz	3.5GHz	12 x 3.5GHz

https://github.com/tesseract-ocr/tesseract/blob/master/ChangeLog

The Dark Ages: 1995-2005: Gathering Dust

By Superbmust (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons
https://commons.wikimedia.org/wiki/File%3APGC.jpg

The Dark Ages

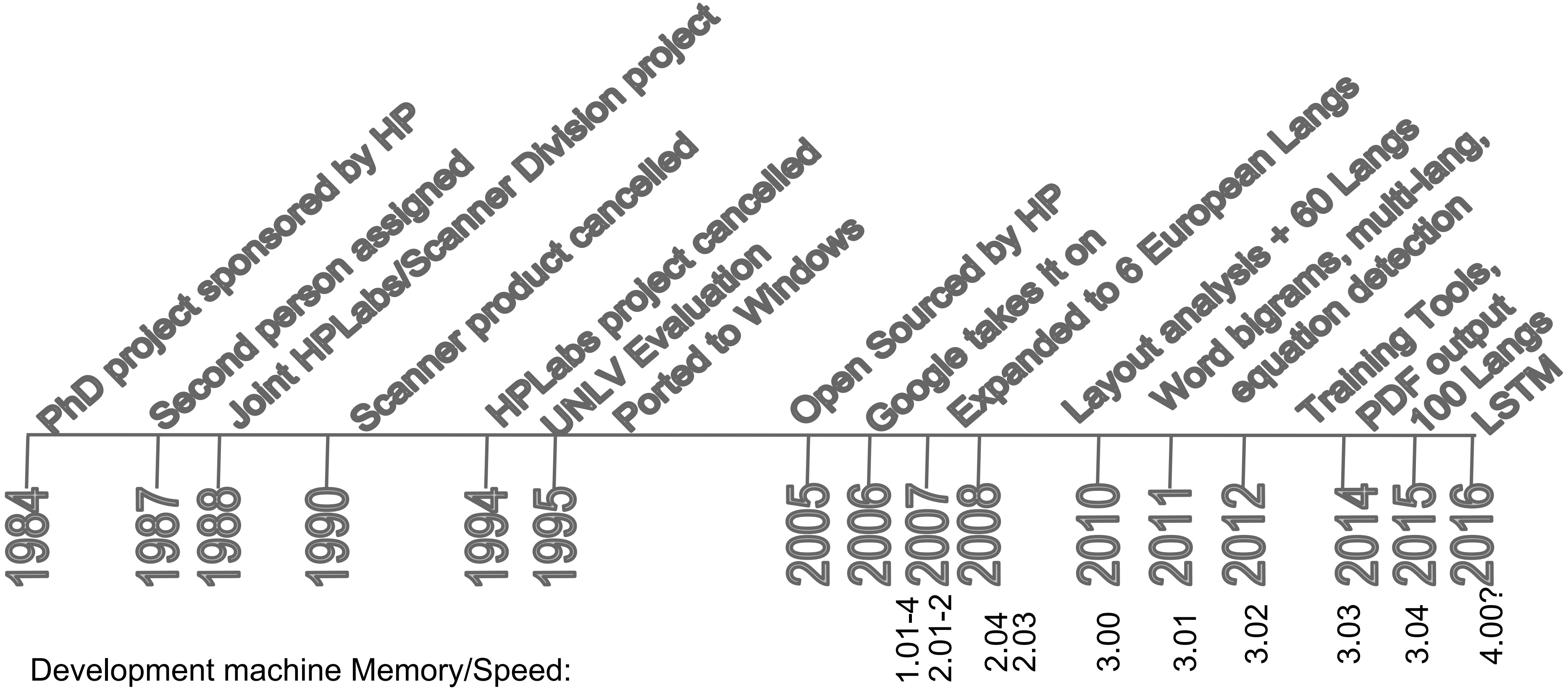
Tesseract:

- Ported to Windows.
- Moore's law:
  - 20x Memory Capacity.
  - 20x Speed.

Commercial Engines:

- 1994: Caere buys Calera.
- 1995: First use of character n-grams in commercial system. (Wordscan 3.1).
- 1996: Caere buys Recognita.
- 2000: Scansoft buys Caere.
- Voting improves accuracy, but engine is much *slower*.

Tesseract Timeline



PhD project sponsored by HP
Second person assigned
Joint HPLabs/Scanner Division project
Scanner product cancelled
HPLabs project cancelled
UNLV Evaluation
Ported to Windows
Open Sourced by HP
Google takes it on
Expanded to 6 European Langs
Layout analysis + 60 Langs
Word bigrams, multi-lang,
equation detection
Training Tools,
PDF output
100 Langs
LSTM

1984
1987
1988
1990
1994
1995
2005
2006
2007
2008
2010
2011
2012
2014
2015
2016

1.01-4
2.01-2
2.04
2.03
3.00
3.01
3.02
3.03
3.04
4.00?

Development machine Memory/Speed:

4MB	16MB	64MB	256MB	3GB	24GB
10MHz	25MHz	100MHz	166MHz	3.5GHz	12 x 3.5GHz

https://github.com/tesseract-ocr/tesseract/blob/master/ChangeLog

2005: HP Open Sources Tesseract

The Open Source Era: 2006 - Present

Tesseract source is available from:

https://github.com/tesseract-ocr/

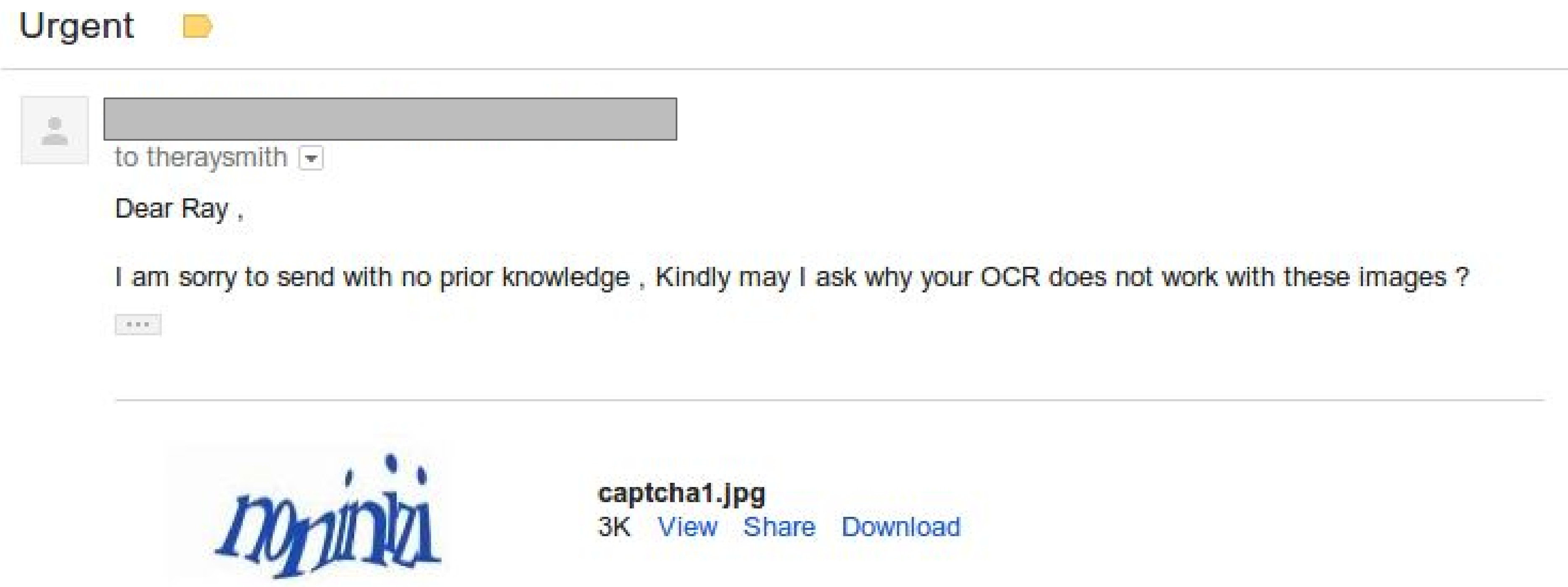
and for direct install on Linux via:

```
apt-get install tesseract-ocr
```

~100 Languages available.

Open Source OCR is Used for Various Applications

From my inbox:



Urgent

[redacted]

to theraysmith

Dear Ray ,

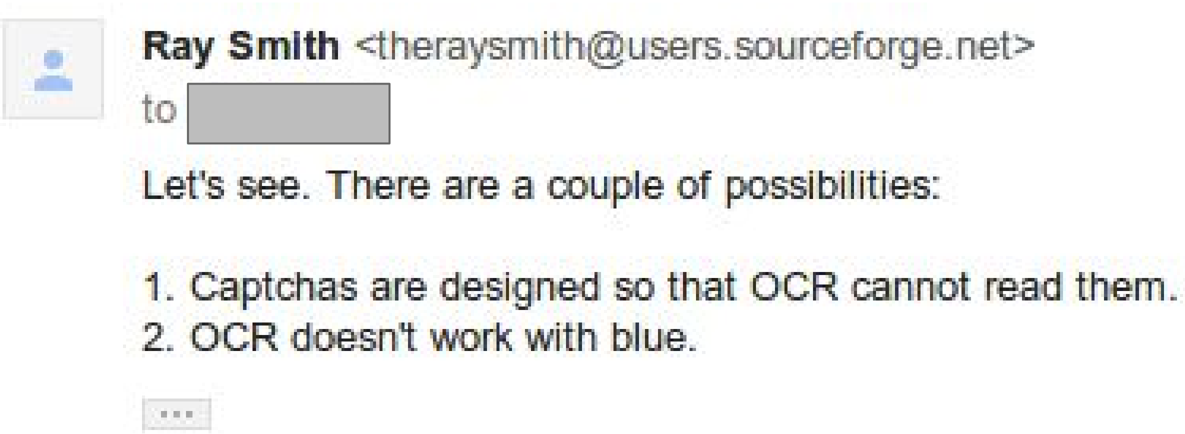
I am sorry to send with no prior knowledge , Kindly may I ask why your OCR does not work with these images ?

captcha1.jpg

3K

Open Source OCR is Used for Various Applications

My reply:



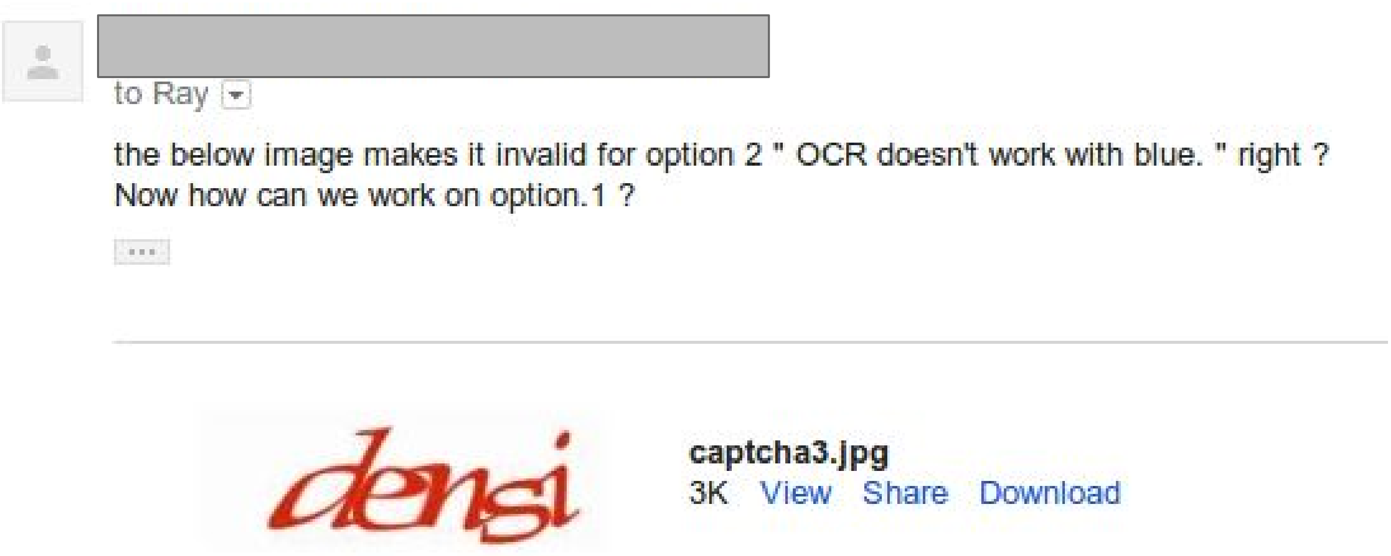
Ray Smith <theraysmith@users.sourceforge.net>
to [redacted]

Let's see. There are a couple of possibilities:

1. Captchas are designed so that OCR cannot read them.
2. OCR doesn't work with blue.

Open Source OCR is Used for Various Applications

He doesn't give up that easily:



[redacted]

to Ray

the below image makes it invalid for option 2 " OCR doesn't work with blue. " right ?
Now how can we work on option.1 ?

captcha3.jpg
3K View Share Download

Recent Improvements

- Internationalized to 100+ Languages.
- Add Full Layout Analysis.
- Table Detection.
- Equation Detection.
- Better Language Models.
- Improved Segmentation Search.
- Word Bigrams.
- PDF Output
- Training Tools

Thanks for Listening!

Questions?