HONG KONG
OPEN SOURCE
CONFERENCE
2017

Open-Source Chinese Handwriting Recognition (Lightning Talk for Hong Kong Open Source Conference 2017)

Cheung Wai Ho, Chris

chrischeungnf@gmail.com

June 10, 2017

Self introduction

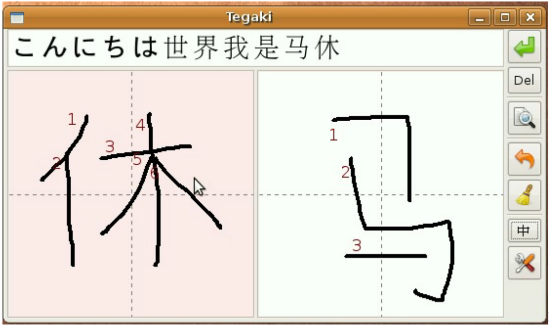
- MPhil in Economics, Year 1 student @ CUHK
- Like programming and the concept of open source
- Part-time teaching assistant
- Assisted teaching of Python for UG students majoring in Economics @ CUHK

Contact information

- GitHub: https://github.com/chrischeungnf
- Facebook: https://www.facebook.com/chrischeungnf
- Blogger: http://chrischeungnf.blogspot.hk
- LinkedIn: https://www.linkedin.com/in/chrischeungnf

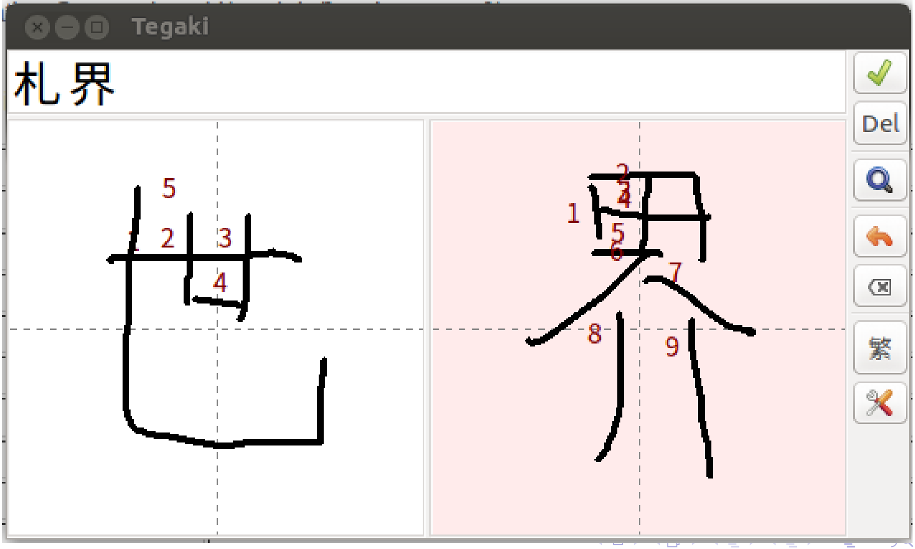
Discovery of Tegaki Project

- Under Linux, I mostly use ibus-table-cangjie3 to type Chinese
- But, sometimes, I do not know how to type some Chinese characters in cangjie3
- Only when I can connect to the Internet, I can use http://hanzi.unihan.com.cn/Qpen instead
- Tegaki [1] is an open-source, local Chinese and Japanese handwriting recognition system

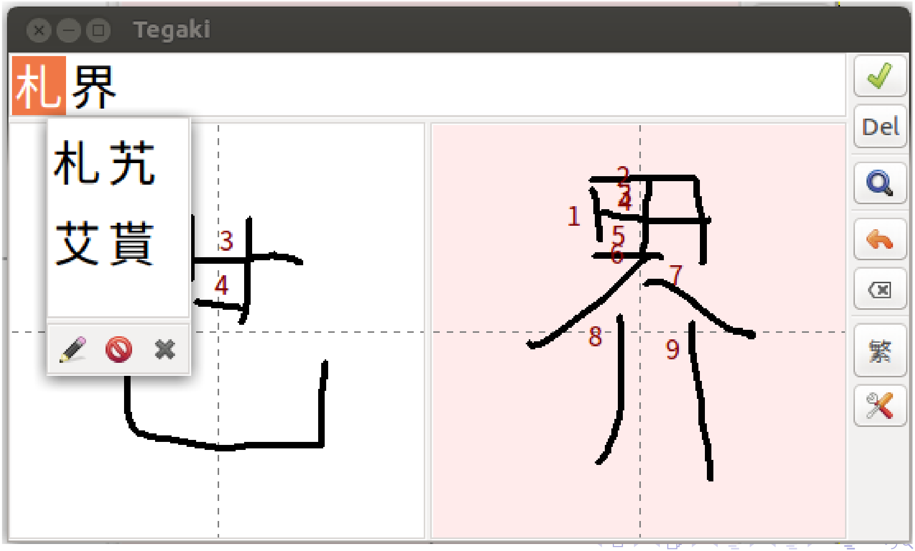


こんにちは世界我是马休

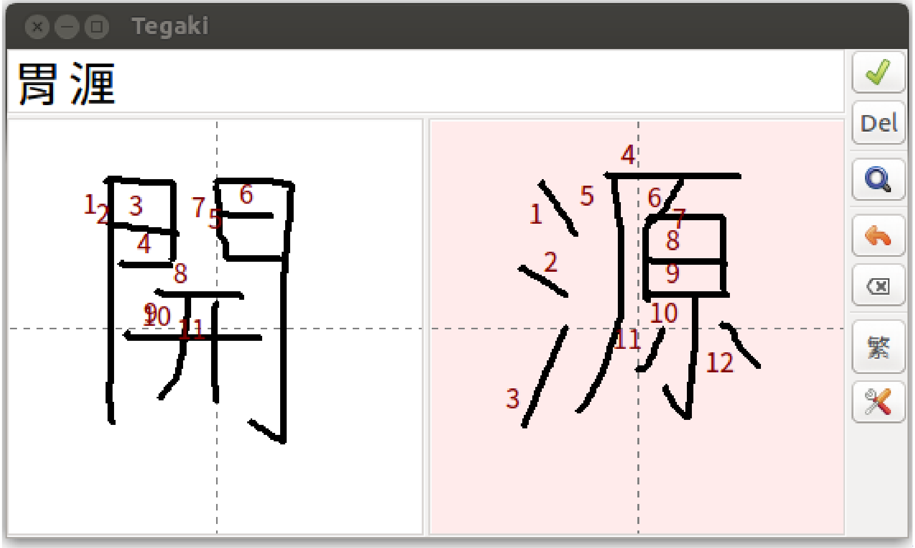
Problem: Bad Chinese (traditional) handwriting recognition I



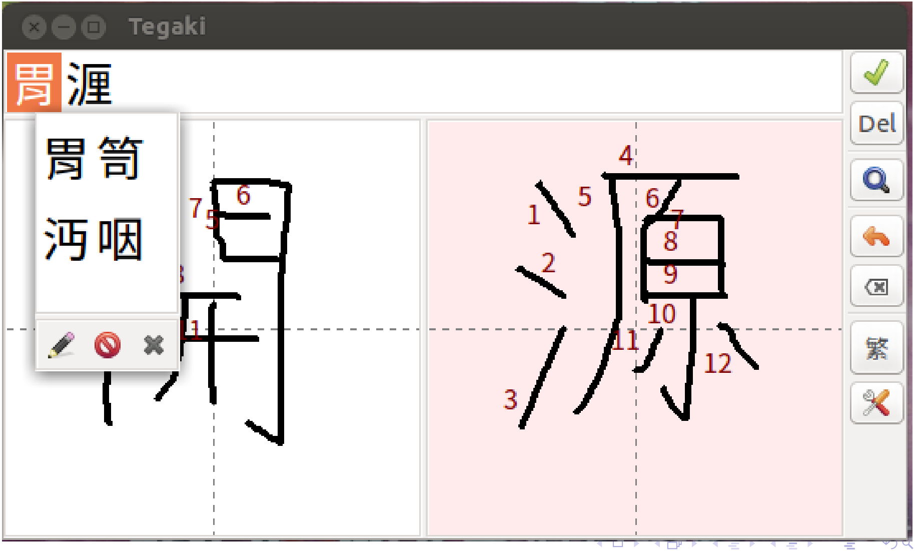
Problem: Bad Chinese (traditional) handwriting recognition II



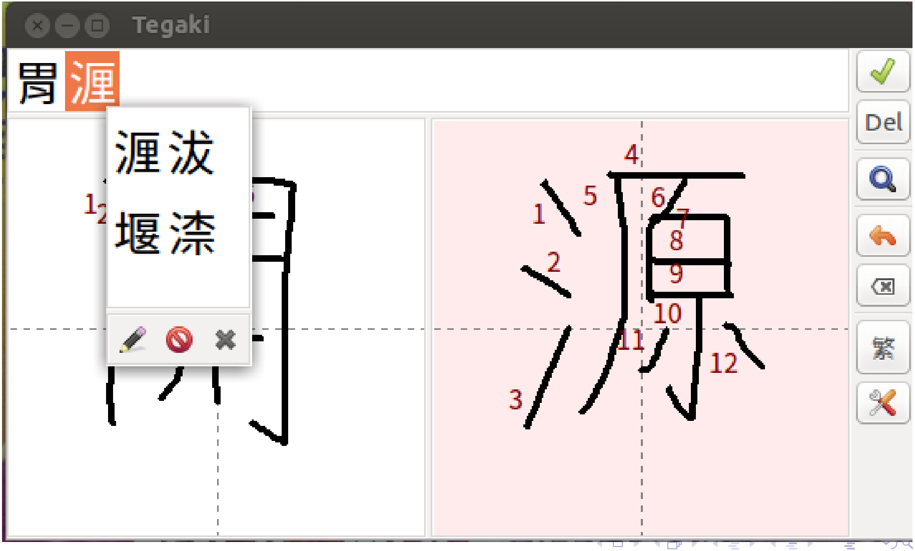
Problem: Bad Chinese (traditional) handwriting recognition III



Problem: Bad Chinese (traditional) handwriting recognition IV



Problem: Bad Chinese (traditional) handwriting recognition V

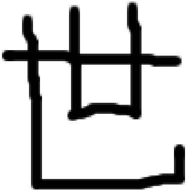


Solution: Use tesseract

- Traditional Chinese model (Zinnia engine) available at https://github.com/tegaki/tegaki/releases/download/v0.3/tegaki-zinnia-traditional-chinese-0.3.zip is used
- The model seems to use rare Chinese characters for training, resulting in poor recognition
- Tesseract OCR [2] seems to be much better

Testing tesseract I

Input:



Output:

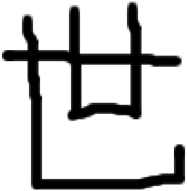
●

Command:

```
tesseract world.png out -l chi_tra -psm 10
```

Testing tesseract II

Input:



世

Output:

世

Command:

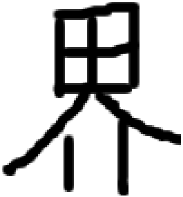
```
tesseract world.png out -l chi_tra -psm 6
```

Tesseract: psm

-psm N
Set Tesseract to only run a subset of layout analysis and assume a certain form of image. The options for N are:
0 = Orientation and script detection (OSD) only.
1 = Automatic page segmentation with OSD.
2 = Automatic page segmentation, but no OSD, or OCR.
3 = Fully automatic page segmentation, but no OSD. (Default)
4 = Assume a single column of text of variable sizes.
5 = Assume a single uniform block of vertically aligned text.
6 = Assume a single uniform block of text.
7 = Treat the image as a single text line.
8 = Treat the image as a single word.
9 = Treat the image as a single word in a circle.
10 = Treat the image as a single character.

Testing tesseract III

Input:



Output:

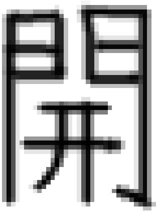
界

Command:

```
tesseract world2.png out -l chi_tra -psm 10
```

Testing tesseract IV

Input:



Output:

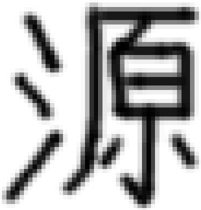
開

Command:

```
tesseract openhw2.png out -l chi_tra -psm 10
```

Testing tesseract V

Input:



Output:

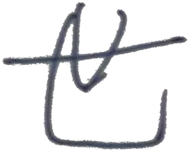
源

Command:

```
tesseract sourcehw.png out -l chi_tra -psm 10
```

Testing tesseract VI

Input:



Output:

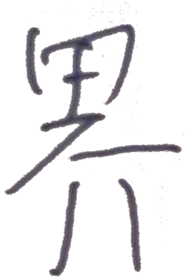
芭

Command:

```
tesseract worldrhwo.png out -l chi_tra -psm 10
```

Testing tesseract VII

Input:



Output:

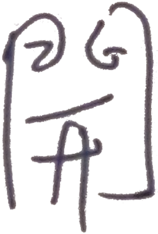
縉

Command:

```
tesseract world2rhw.png out -l chi_tra -psm 10
```

Testing tesseract VIII

Input:



Output:

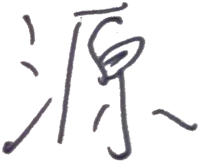
闌

Command:

```
tesseract openrhw.png out -l chi_tra -psm 10
```

Testing tesseract IX

Input:



Output:

腮

Command:

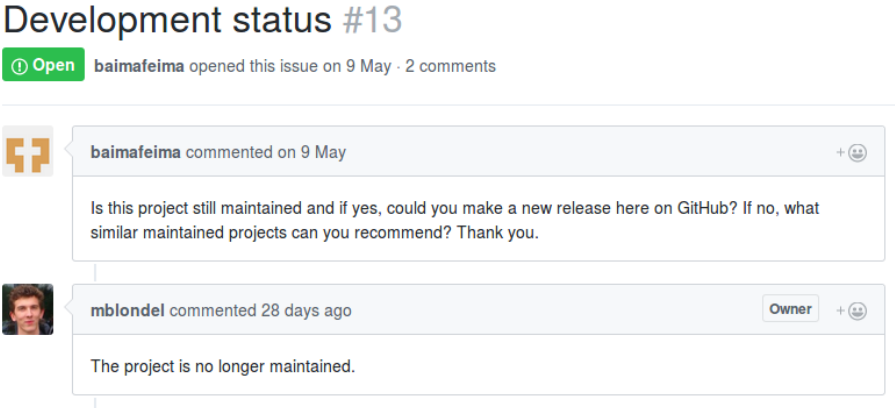
```
tesseract sourcerhw.png out -l chi_tra -psm 10
```

A lesson

Think about alternatives

My plan

- Improve tesseract's Chinese (traditional) handwriting recognition
- Integrate tesseract into tegaki?



Development status #13

Open baimafeima opened this issue on 9 May · 2 comments

baimafeima commented on 9 May

Is this project still maintained and if yes, could you make a new release here on GitHub? If no, what similar maintained projects can you recommend? Thank you.

mblondel commented 28 days ago

Owner

The project is no longer maintained.

Joining this project

- My project is at https://github.com/chrischeungnf/tegaki-traditional-chinese-local

Reference

Tegaki Project
Tegaki - Open-Source Chinese and Japanese Handwriting Recognition
Retrieved from: https://tegaki.github.io/

Tesseract OCR
Available at: https://github.com/tesseract-ocr/tesseract