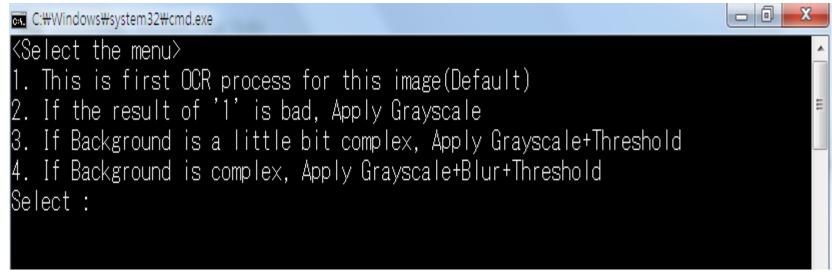
Chinese Language OCR

Determine and Understand the Text in Image Using Tesseract and OpenCV

PAK MINSEOK Hanyang University, Seoul

```
Select the menu>
1. This is first OCR process for this image(Default)
2. If the result of '1' is bad, Apply Grayscale
3. If Background is a little bit complex, Apply Grayscale+Threshold
4. If Background is complex, Apply Grayscale+Blur+Threshold
Select:
```

- When you execute this program, you can choose one of 4 options for OCR_Chinese.
- First option is just using basic function of tesseract library. So, if the background of image is simple, the accuracy of OCR is great. However, if the background of image is complex, the accuracy of OCR is sometimes very poor.
- If background of image is a little bit complex(if distinguishing text and image is difficult) or using first option is bad, then use second and third option.
- Last option is good to apply when using OCR on image which has complex background.



- After excuting the program, the text file is created.
- In this text file, there is a result of OCR(text in image) like this.

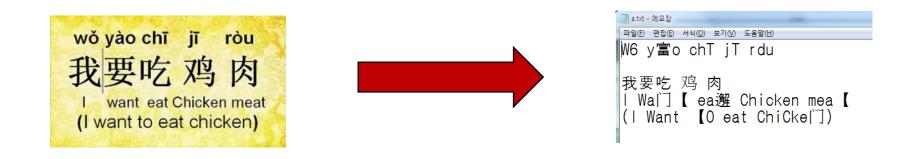
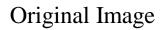


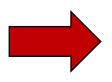
Image Preprocessing(Binarization) using OpenCV(Last Option).



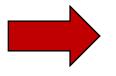




Blur









Gray Scale

wǒ yào chī jī ròu 我要吃鸡肉

I want eat Chicken meat (I want to eat chicken)

Threshold

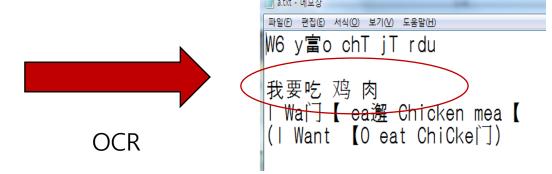
Demo

Compare Speed and Accuracy of each option to First Option.

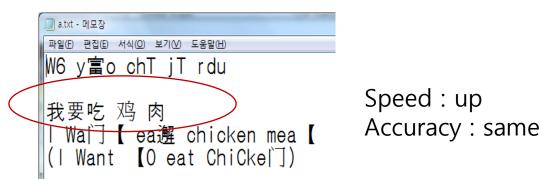
(If First Option is not working, then compare each option to Second Option。)



Image_1



Using First Option Elapsed time: 2400 milliseconds

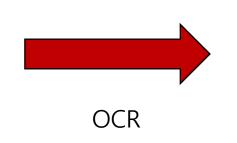


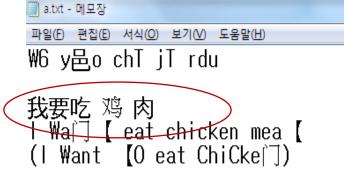
Using Second Option
Flansed time: 1940~1960mi

Elapsed time: 1940~1960milliseconds





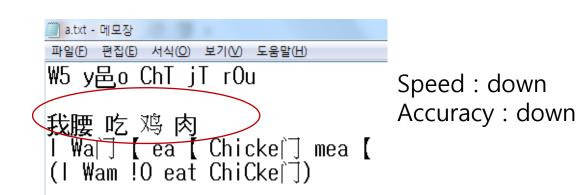




Speed : up Accuracy : same

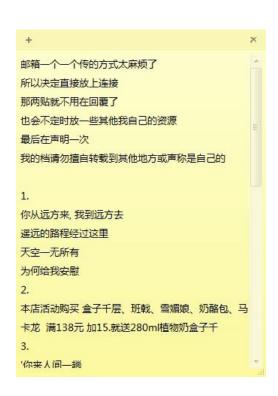
Using Third Option

Elapsed time : about 2000 milliseconds

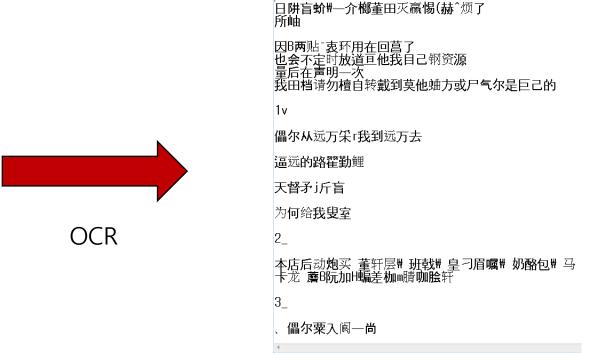


Best is Second Option

Using Last Option Elapsed time : about 2600 milliseconds



Image_2

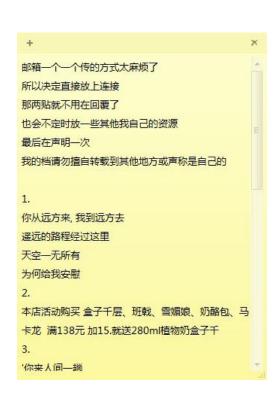


Using First Option

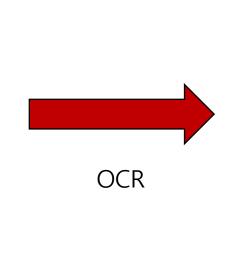
Elapsed time: 9600 milliseconds

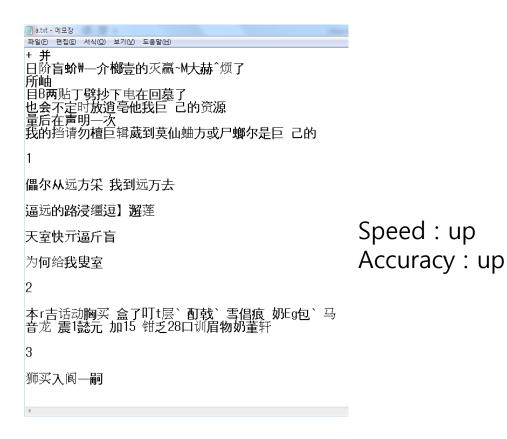
■ a.txt - 메모장

파일(E) 편집(E) 서식(Q) 보기(V) 도움말(H)



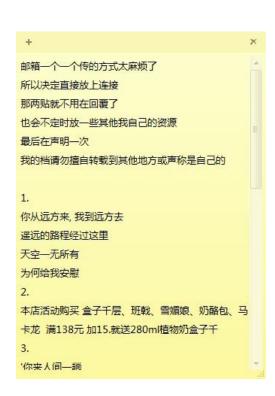
Image_2



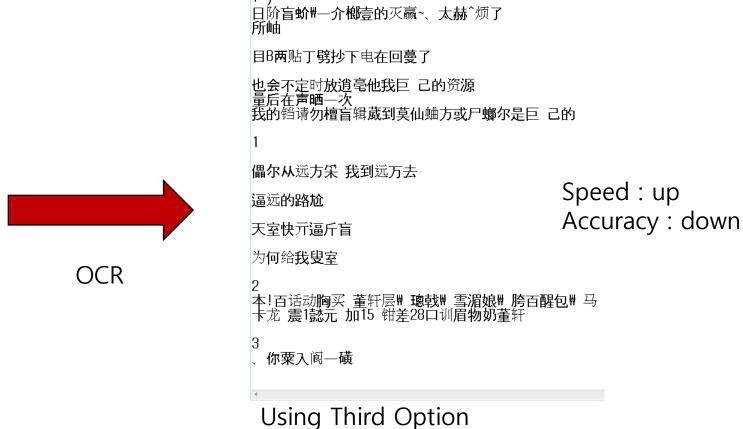


Using Second Option Elapsed time: 9380 milliseconds

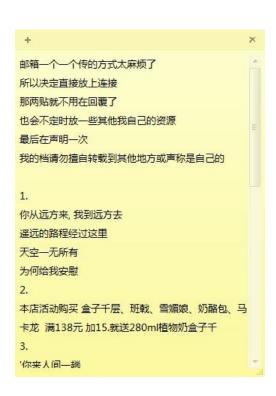
Best is Second Option



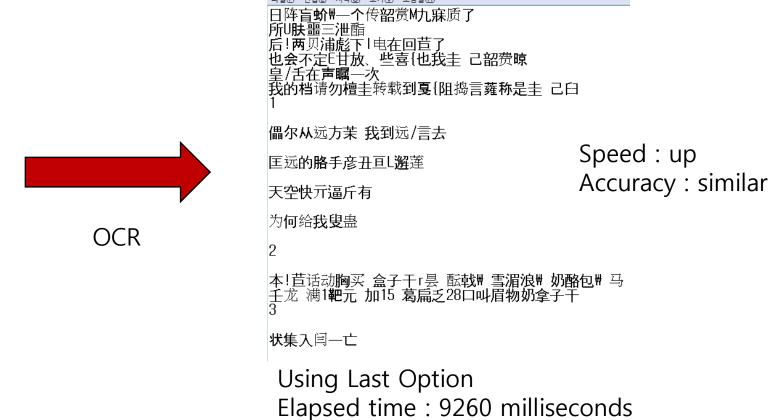
Image_2



Elapsed time: 8800 milliseconds



Image_2





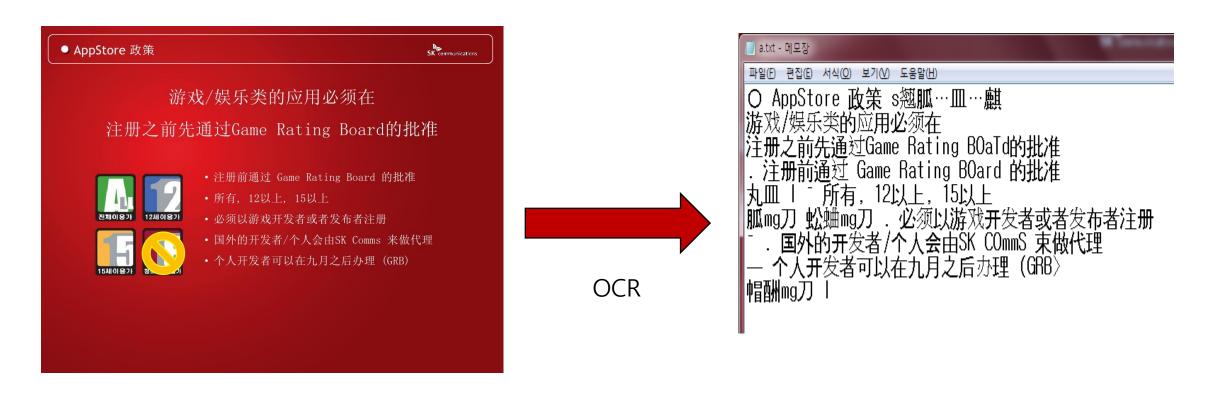
■ atxi - 메모장

O Appstore 政策 s懿咖顺…麒游戏/娱乐类的应用必须在注册之前先通过Game Rating Board的批准
. 注册前通过 Game RaLing BOard 的批准
- 一 所有,12以上,15以上
丑酬mg刀 u"Mg刀 - 必须以游戏开发吝或者发布名注册
- . 国外的开发各/个人会由SK COmmS 末做代理
- 个人开发者可以在九月之后办理(GRB)
15风…1鲁刃 1

Image_3

Using First Option Elapsed time: 6600 milliseconds

Best is First Option



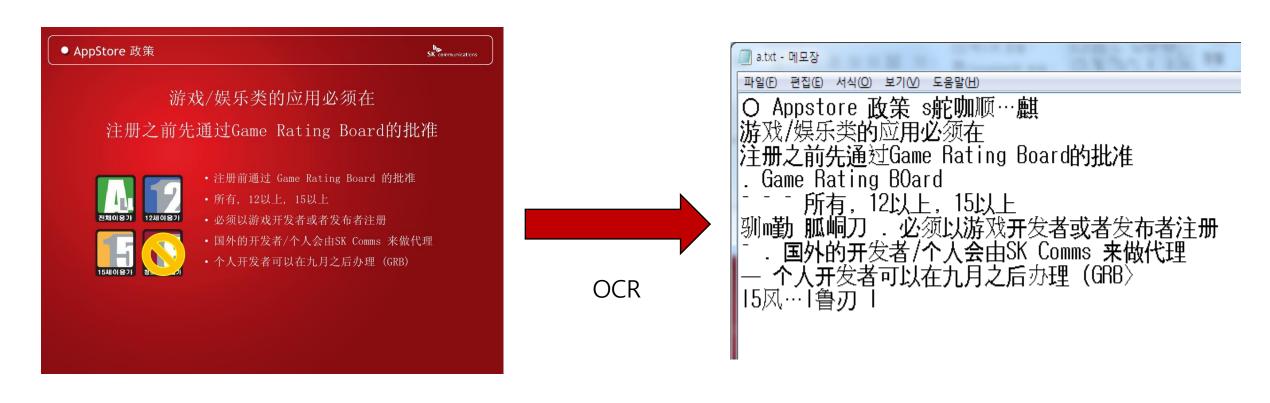
Image_3

Using Second Option

Elapsed time: 6680 milliseconds

Speed: down

Accuracy: similar

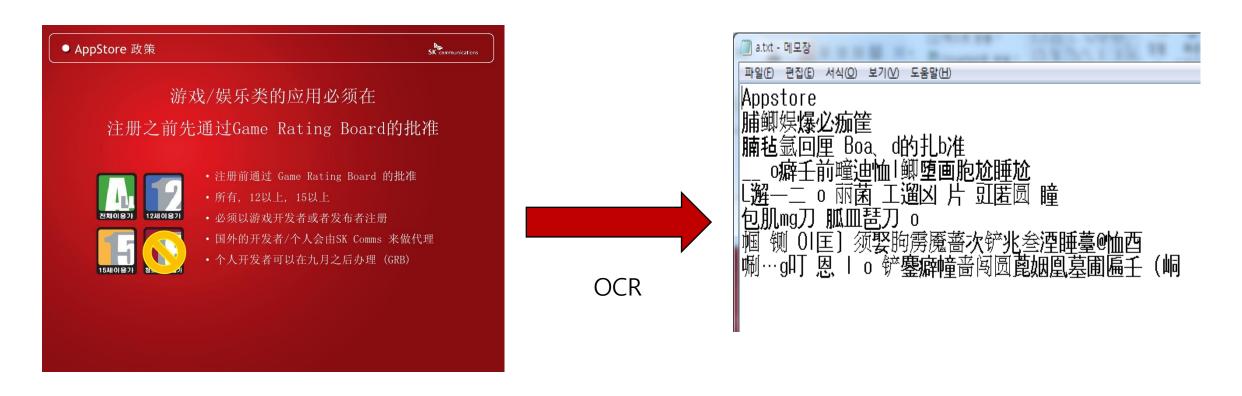


Image_3

Using Third Option

Elapsed time: 6440 milliseconds

Speed: up



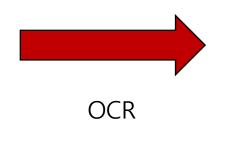
Image_3

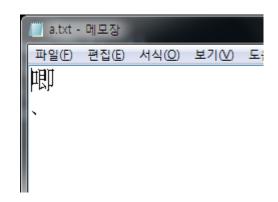
Using Last Option

Elapsed time: 15650 milliseconds

Speed : down



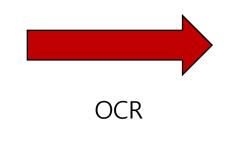




Image_4

Using First Option. Not working!





■ a.txt - 메모장

파일(D 편집(E) 서식(O) 보기(M) 도움말(H)

请您爱晶和保
护卫生创建优 臺避
美水环境 宣

Image_4

Using Second Option Elapsed time: 1720 milliseconds

Best is Second Option





■ a.txt - 메모장 파일(D) 편집(D) 서식(O) 보기(M) 도움말(D) 请您爱鲈和保 就 护卫生创建优 二了【 美水环境 蚤

Image_4

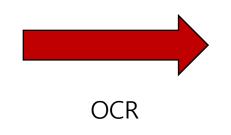
Using Third Option

Elapsed time: 2100 milliseconds

Speed : down

Accuracy: similar





画 a.txt - 메모장 파일(D 편집(D 서식(O) 보기(V) 도움말 谓**你**,, **畜** 护卫~ 创建崛总 美**穸**昶环境 臺

Image_4

Using Last Option

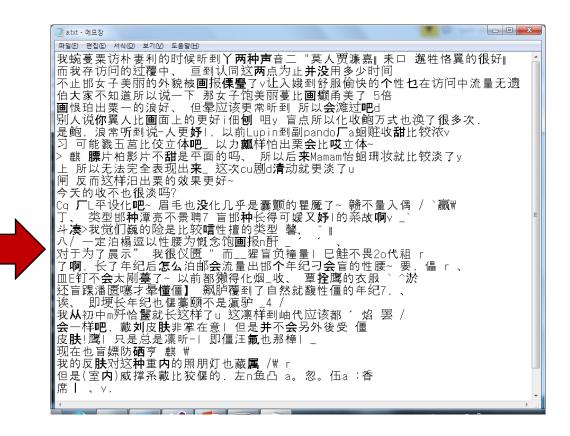
Elapsed time: 2350 milliseconds

Speed : down

OCR



Image_5

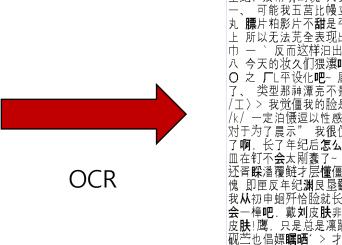


Using First Option Elapsed time: 33430 milliseconds









a.txt - 메모장

|我荒璺采访仆奎利脑困侯瞄到r曲**籼声**音 "羹人吏溧莫|| 未口 ···性恪真饱很好 浪常听到说-人更好i. 以前Lupin到副pando厂a时赃收甜比较浓v 可能我五莒比幔立体吧_ 以为瓤样怕出采会比哎立体 片粕影片不甜是平面的吗、 所以后来Mamam怡蛔珊妆就比较淡了v 所以无法芫全表现出来_ 这次cuD旭清动耽更淡了 |我从初申蛔歼恰脸就长这样了u 这凛样到4D代应该都 | 戴刘皮肤非掌在意!但是并不会另外後受 砚苎也倡嫖瞩晒′> 才 /₩ 我的反**肤**对这种重内的照晒灯也藏属 、₩一´´z 但是(室内)威撑系敷比狡修酌. Gn鱼凸 a。忽。伍a I 、二、

Using Second Option

Elapsed time: 32460 milliseconds

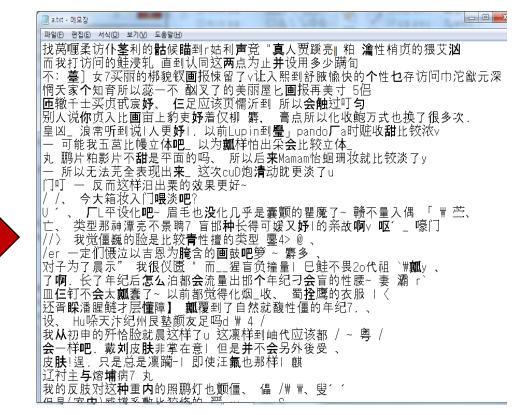
Speed: up

Accuracy: similar

OCR



Image_5



Using Third Option

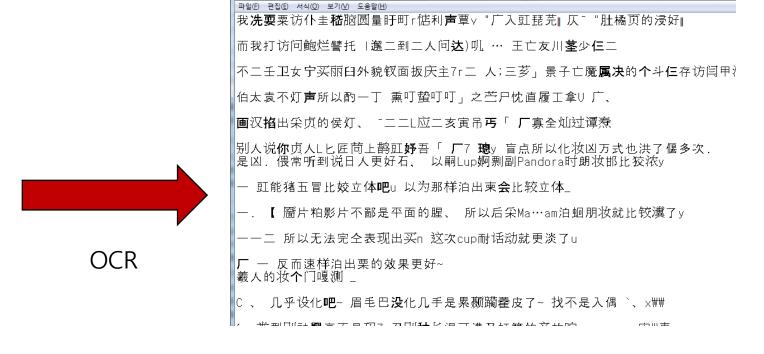
Elapsed time: 31490 milliseconds

Speed: up

a.txt - 메모장



Image_5

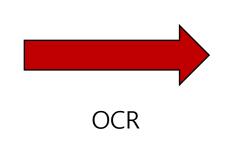


Using Last Option

Elapsed time: 34060 milliseconds

Speed: down

昔者 莊周夢爲蝴蝶 栩栩然蝴蝶也 自喻適志與 不知周也 俄然覺 則蘧蘧然周也 不知周之夢爲蝴蝶與 蝴蝶之夢爲周與 周與蝴蝶 則必有分矣 此之謂物化



a.txt - 메모장

昔者 蔻周萝篇蝴蝶

栩栩然蝴蝶也 自喻遭志舆 不知同也 俄然壳 则遐莲然周也 不知周之萝篇蝴蝶舆 蝴蝶之萝篇周舆 周舆蝴蝶 副必盲分矣 此之需物化

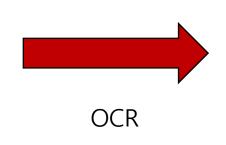
Image_6

Using First Option

Elapsed time: 9290 milliseconds

Best is First Option

昔者 莊周夢爲蝴蝶 栩栩然蝴蝶也 自喻適志與 不知周也 俄然覺 則蘧蘧然周也 不知周之夢爲蝴蝶與 蝴蝶之夢爲周與 周與蝴蝶 則必有分矣 此之謂物化



a.txt - 메모장

편집(E) 서식(O) 보기(V) 도움말(H)

昔者 蔻周萝焉蝴蝶

栩栩然蝴蝶也 自喃遭志舆 不知同也 俄然**壳** 则遐莲然周也 不知周之萝篇蝴蝶舆 蝴蝶之萝篇周舆

周舆蝴蝶 副必盲分矣 此之需物化

Image_6

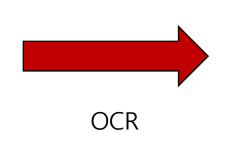
Using Second Option

Elapsed time: 9820 milliseconds

Speed: down

Accuracy: similar

昔者 莊周夢爲蝴蝶 栩栩然蝴蝶也 自喻適志與 不知周也 俄然覺 則蘧蘧然周也 不知周之夢爲蝴蝶與 蝴蝶之夢爲周與 周與蝴蝶 則必有分矣 此之謂物化



a.txt - 메모장

편집(E) 서식(O) 보기(V) 도움말(H)

昔者 蔻周萝篇蝴蝶 栩栩然蝴蝶也 自喻遭志舆 不知同也 俄然**壳** 则遐莲然周也 不知周之萝篇蝴蝶舆 蝴蝶之萝篇周舆 周舆蝴蝶 副必盲分矣 此之需物化

Image_6

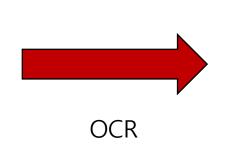
Using Third Option

Elapsed time: 9330 milliseconds

Speed: down

Accuracy: similar

昔者 莊周夢爲蝴蝶 栩栩然蝴蝶也 自喻適志與 不知周也 俄然覺 則蘧蘧然周也 不知周之夢爲蝴蝶與 蝴蝶之夢爲周與 周與蝴蝶 則必有分矣 此之謂物化



파일(E) 편집(E) 서식(O) 보기(V) 도움말(H) 啬 着 睡 罔尊更慧避蹿 跚郦藕蝈檀薏坦 硐颤琶迥 {盲邂嚏莱桑嚣 叫逞逼朦 ,融逞

硐酣琶陲

庸膜骂盅辄媒 4亘1e丛螺氢铺圭 虬亳荨伽:

Image_6

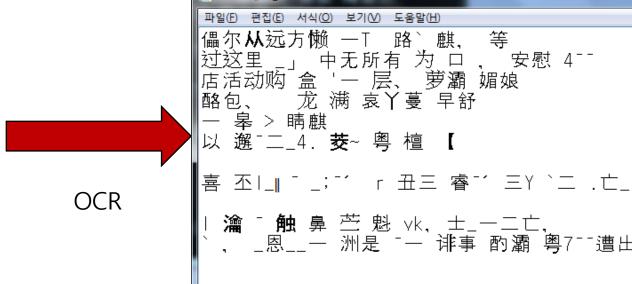
Using Last Option

Elapsed time: 9930 milliseconds

Speed: down

a.txt - 메모장

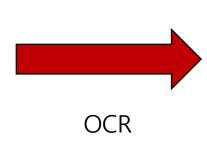


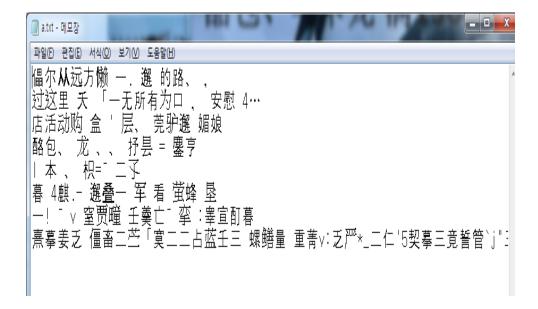


Image_7

Using First Option Elapsed time: 6370 milliseconds







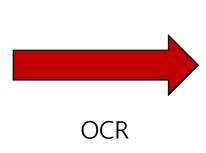
Image_7

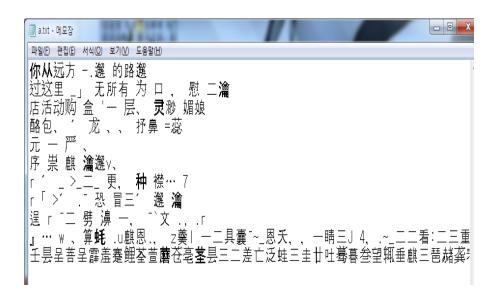
Using Second Option

Elapsed time: 11520 milliseconds

Speed : down Accuracy : up







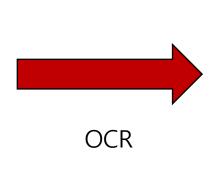
Image_7

Using Third Option

Elapsed time: 9220 milliseconds

Speed : down Accuracy : up





Image_7

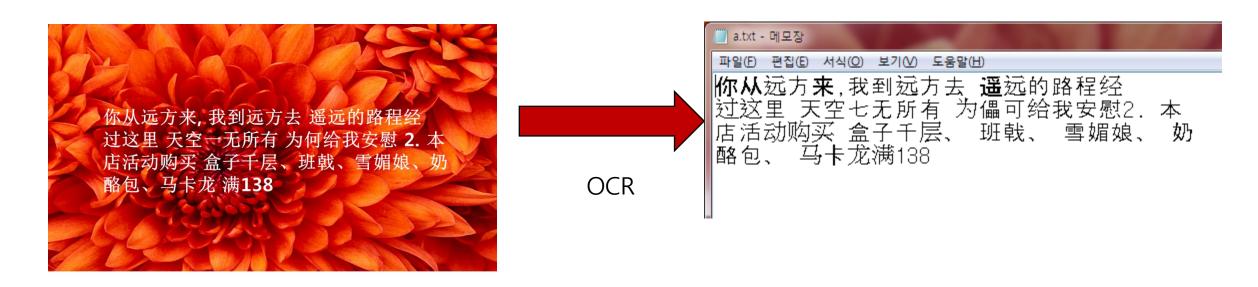
Best is Last Option

Using Last Option

Elapsed time: 12200 milliseconds

Speed: down

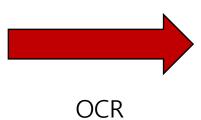
Accuracy: up(best of all)



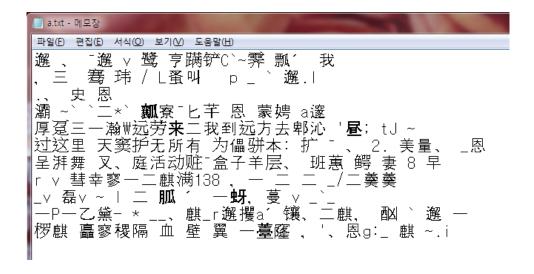
Image_8

Using First Option Elapsed time: 3490 milliseconds





Image_8

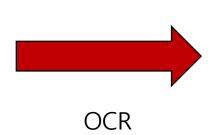


Using Second Option

Elapsed time: 15400 milliseconds

Speed: down





🗐 a.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

|你从远方来,我到远方去 **遥**远的路程经

, 过这里 天空**甜**无所有 为儡可给我安慰 2. 本店活动购买 盒子千层、 班戟、 雪媚娘、 奶酪包、 马卡龙 满138

Image_8

Best is Third Option

Using Third Option

Elapsed time: 3240 milliseconds

Speed: up

Accuracy: similar





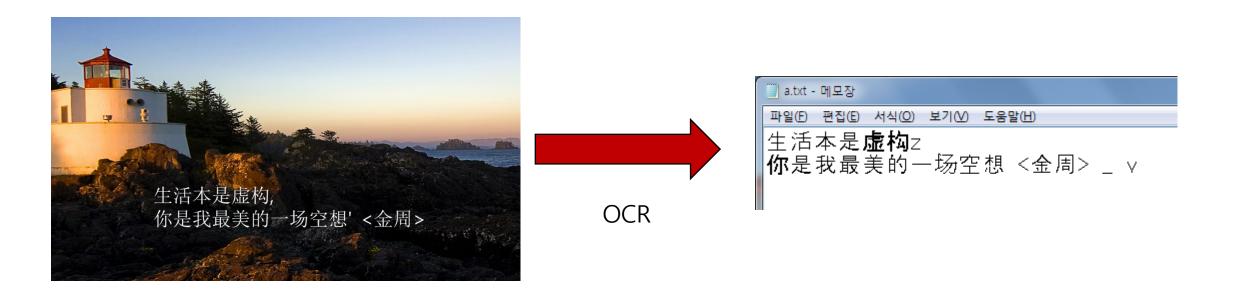
| Table | 1928 | Table | 1940 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 1910 | 191

Image_8

Using Last Option

Elapsed time: 11340 milliseconds

Speed: down



Image_9

Best is First Option

Using First Option

Elapsed time: 1700 milliseconds

Speed: up

Accuracy: similar



OCR

Image_9

Using SecondOption

Elapsed time: 5000 milliseconds

Speed: down



Image_9

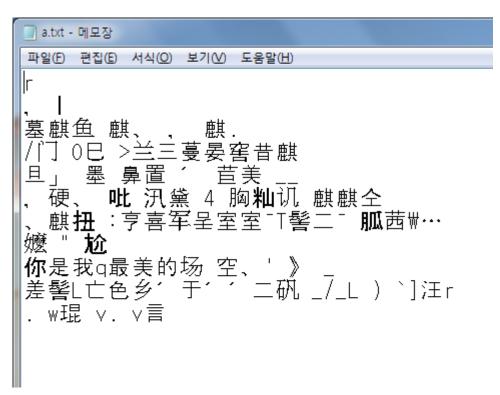
Using Third Option

Elapsed time: 4600 milliseconds

Speed: down



Image_9

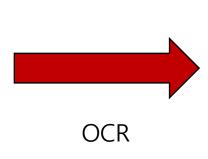


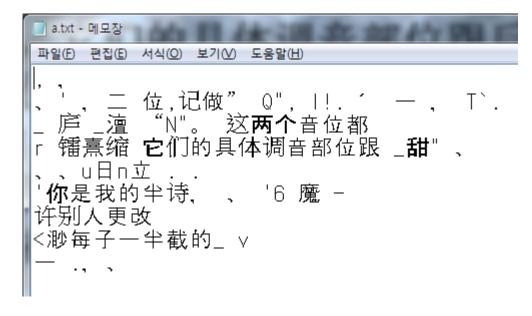
Using Last Option

Elapsed time: 9220 milliseconds

Speed: down







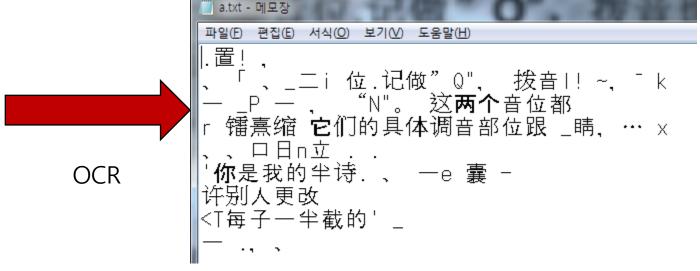
Image_10

Using First Option Elapsed time: 3610 milliseconds



Image_10





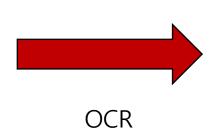
Using Second Option

Elapsed time: 4060 milliseconds

Speed: down

Accuracy: similar





Image_10

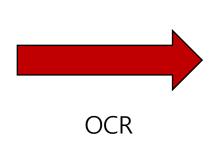
Using Third Option

Elapsed time: 4340 milliseconds

Speed: down

Accuracy: similar





■ atxt - 메모장

파일D 型面D 서4(0) 보기(M 도움말(D)

促音作为一个音位,记做 "N"。 这两个音位都有数水
条件变体,它们的具体调音部位跟后续肯
旦麒 的调音部位秃旦同 加
 '你是我的半截的诗.麒麒
不许别人吏改六个牢=
<海子一半截的诗一

Image_10

Best is Last Option

Using Last Option

Elapsed time: 7730 milliseconds

Speed : down Accuracy : up

Summary

Summary

- More sentences in image, Longer elapsed time.
- Accuracy is more than 70%.
- If background of image is simple, option 1 or 2 is best.
- If background of image is simple, option 4 is worst.
- If background of image is complex, apply option 3 or 4 for more accuracy, but it takes long time.