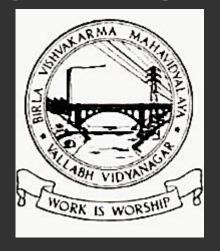
# **BVM Engineering College, VV Nagar**







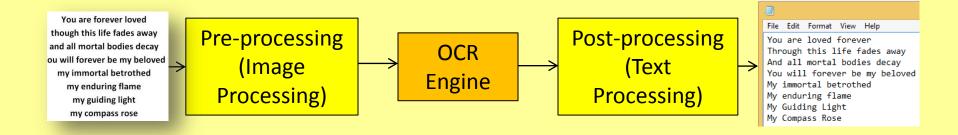
# PROJECT #1

# Electronics & Communication Dept.

# **Presented By:**

- Chaitanya Tejaswi (140080111013)
- Shahnawaz Yusufzai (120080112036)

## **OCR System: Process Diagram**



## **OCR-PIA: Objective**

Modern-day applications integrate diverse functionality and package them in a standalone Android (.apk) file. One such application is an Optical Character Recognition (OCR) module, which interprets text from an input image, and displays it in raw text format. The nature of android makes it difficult for non-Java programmers to work with existing apks. The work to be carried out will thus involve programming an OCR system in a scripting language (such as Python), and implementing it on an Android base. Native functionality will be added as supported by the Kivy framework, which enables building standalone apks using Python scripts.

# **Steps Associated with Pre-Processing**

# **Image processing**

Rescaling

Binarisation (Thresholding)

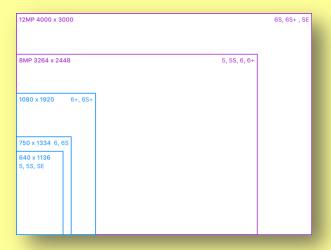
Rotation (De-Skewing)

**Border Removal** 

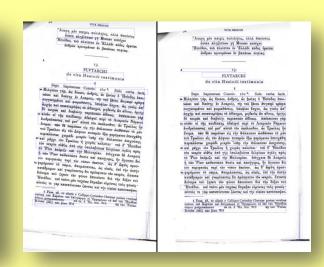
**Noise Removal** 

**Page Segmentation** 

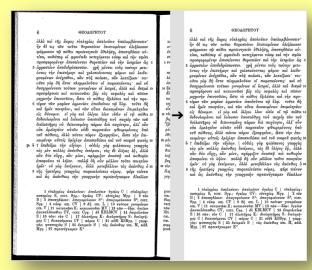
**Dictionaries, Word lists & Patterns** 







Rescaling Binarisation Rotation



**Border Removal** 

88 GEOLASPHTOT

Only, the glatene deslegger; are groundly one veryon; former
6 sty, despite, contained sign deed; pairs appelled so
6 sty, despite, contained sign deed; pairs appelled so
6 sty, despite, contained sign deed; pairs appelled so
6 sty, despite, contained sign deed; pairs appelled so
6 sty, despite, contained sign deed; pairs appelled so
6 sty, despite, contained significant services;
6 sty, despite, despite significant services;
6 sty, despite, despite significant services;
6 sty, despite, despite, significant services;
6 sty, despite, significant despite, despite, despite,
6 sty, despite, significant despite, despite, despite,
6 sty, despite, despite, significant services;
6 sty, despite, despite, significant services;
6 sty, despite, despite, despite, despite, despite,
6 sty, despite, despite, despite, despite, despite,
6 sty, despite, despite, despite, despite, despite,
6 sty, despite, despite, despite, despite,
6 sty, despite, despite, despite,
6 sty, despite, despite, despite,
6 sty, despite, despite,
6 sty, despite,
6 s

**Noise Removal** 

### **OCR: Problem Identification**

Generic OCR Engines assumed the image to be scanned (by a *flatbed scanner*). So, the work was focused on developing robust *OCR Engines*.

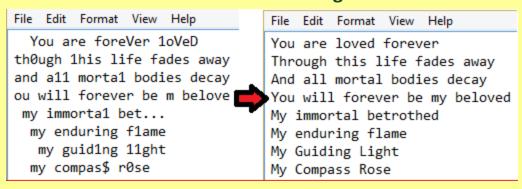
Today, most images are captured using *smartphones*. Implementing an OCR system for such images needs us to reconsider the defects generated in the process. So, the focus should be on effective *Pre-Processing* of images, while making use of existing OCR techniques.

We must also note that OCR Systems are prone to make mistakes, and generate text that might not make much sense. So, additional *Post-Processing* of text will improve efficiency of the process.

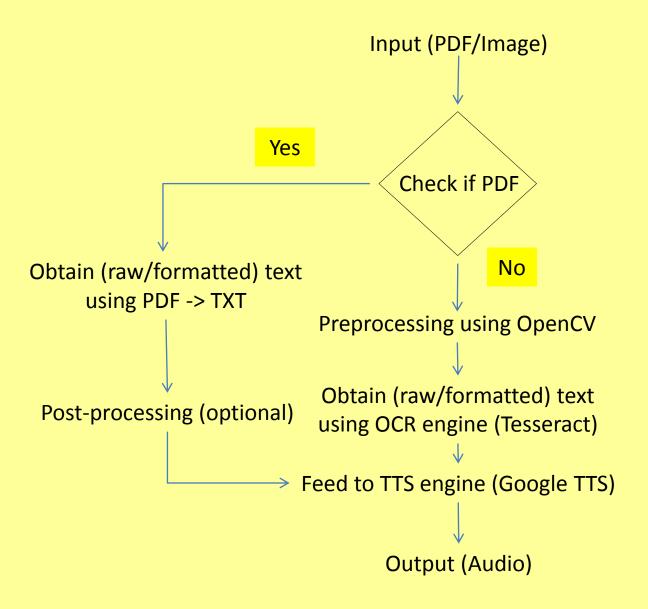
Pre-Processing



### **Post-Processing**



### **OCR-PIA:** Flowchart

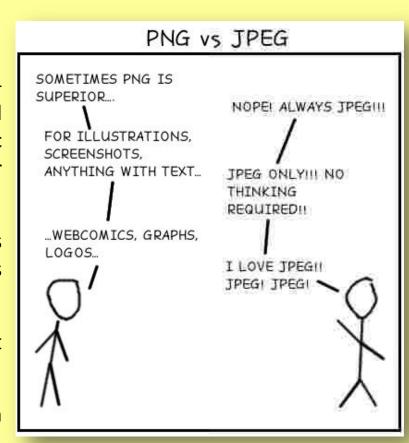


# To JPEG or PNG, that is the question

JPEG uses lossy compression, PNG uses lossless compression.

#### WHAT DOES THIS MEAN?

- JPEG was designed for compressing fullcolour/grayscale images of natural, real world scenes. It works well on photographs or naturalistic artwork, but not so good with lettering, cartoons, or line-drawings.
- Since human eyes perceive small colour-changes less accurately than small brightness-changes, JPEGs are great for us to look at. Plus, it saves space!
- But for OCR, we need high resolution image to start with, since we will lose pixels when processing it.
- PNG, which was meant for the Web, is better than GIF, and less complex than TIFF.
- But most smartphones save camera-generated images as JPEGs.



# JPEG to PNG?

# Is there is a simple algorithm to convert between jpeg and png?



Steve Baker Answered Dec 19, 2016

It's not simple if you try to do it all yourself - but that would be decidedly crazy.

Grab copies of "libpng" and "libjpeg".

Use libjpeg functions to load the jpeg file into memory - the library will happily unpack it into an array of red/green/blue values.

Then call the libpng functions to write that array back out again.

- But since JPEG uses lossy compression, and PNG uses lossless, it's necessary to know what goes on in the conversion process.
- WHY? Just to be sure that we're actually benefitting from the conversion process, and not wasting any useful pixel-data.

## **Some Results**









We performed binarisation using *Global-Thresholding* methods in *OpenCV*. But the time for conversion using *Tesseract-OCR* was found to be non-trivial. The results for *Adaptive-Thresholding* methods didn't yield any significant improvement. However, *Otsu's Algorithm* did provide quicker outputs in certain cases.

#### Original Image



### TRUNC



Original Image



TRUNC



BINARY



**TOZERO** 



**BINARY** 



TOZERO



BINARY INV



TOZERO\_INV



BINARY INV



TOZERO\_INV



Original Image



TRUNC



Original Image





**BINARY** 



TOZERO



**BINARY** 



**TOZERO** 



BINARY INV



TOZERO\_INV



BINARY INV



TOZERO INV



- a × Python 3.6.1 Shell THRESH\_TRUNC\_1 File Edit Shell Debug Options Window Help ----- RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ------Image to be processed: lenovo\_k6power\_1.jpg 45 Seconds drive, but kept its cables and charger? You are not alone. .' Just look in any household junk d raWer w and you 11 nd cords that have come and gone 'relics of the camcorder era, laptops and di gi- tal cameras past, and outdated phone (10115 gles, says an article in The Verge. - ~ 6 H 1. We can t bring ourselves to part With them because we are not sure we Would not need them again. P hones have become efcient and their batteries larger, yet leaving home with- oat a Charger makes us anxious. . A battery, no matter how sophisticated, 1s eeting. When we have our cords With us, w e are in constant pursuit Of power, even When we are fully charged, as a form Of security We often discover our misfortune - the loss of power - When lt 3 too late. The opposite of \_ A being f ully charged 13 dead. .. T Will our lives ever be Wire- -free again? Maybe not.Apple introdu ced wireless ear . phones last year, and now Dell has a Wireless ~ 5; chargmg laptop, but electric ears and their ===== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ======= 38 Seconds Image to be processed: lenovo\_k6power\_la.jpg C! ,- 1. . . 11 ' / .1 . 1- . H 1; , , ". 1 . . . uv I  $\,$  V  $\,$  V , . . . J. . , ,  $\,$  - y  $\,$  a . . - '  $\,$  V 'VK . 7 .  $\,$  x  $\,$  ave you ever thrown away a dead gadg , 1 - .7  $\,$  et, say an old tablet or an eXternal hard, 7 drive, but kept its cables and charger? You are not alone. 4 Just look 1n any hou sehold junk drawer 1 and you 11 nd cords that have come and gone ' relics of the camcorder era, laptops and digi- tal cameras past and outdated phone don- glee, says an article in The Verge. W can t bring ourselves to part With them " because we are not sure we Would not need .. . them ag aln. Phones have become efcient and I ) their batteries larger, vet leaving home With out a Charg er makes us anxious. . A battery, no matter how soph1sticated,1s eeting When we have our cords Wi th us, we are in constant pursuit of power, even When we are fully charged, as a form Of security We often discover our misfortune ~ the loss of power - When 1t 3 too late. The opposite of bei ng fully charged IS dead. ' , 1 . Will our lives ever be Wire- -free again? Maybe not Apple in troduced Wireless ear- phones last year, and now Dell has a Wireless 3 Charging laptop, but electric cars and their 1 'chargers are only beginning to arrive.

- Ø × Python 3.6.1 Shell THRESH TRUNC 3 File Edit Shell Debug Options Window Help Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 64 bit (AMD64)] on win32 Type "copyright", "credits" or "license()" for more information. ----- RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ------24 seconds Image to be processed: 3a.jpg "mm mm \_ , " 'ItI?1'mt!13'); iili;f4p;\$4 g.3'83 Md mr (1.90 r 1119 Wnuv. in applmim king in 110111 of fowes 9113111 gal m the , in susmlned comma geomb term Opel mans m I li on I . Ionda3 honoured two L RPF mked commlmldants with Kim a mob S Chakm for a cts of bravery in under J K and conferred 40 of the eIndi- 190 police gallanm' medals a the E anar ded this Independence ham. 5 Day on J 5.1K police and Atro- CRPF1B8F personnel deplo- admyed inthe allex At the same Din-S t1mga2pezsonnel of Andhra ----- RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ------72 seconds Image to be processed: 3b.jpg 1",3331EMH 9111111 1; 111M136 m1 t3133-33 33313 3.3!; 113111 1.111111 11131111 111113.131 131119 two 1 RFF 01111111141 3 11111213111113111 x 1111 11111 . 5 111121 E L 11 1311'11'0 31115 11111
'1111'111'1111 Lmder 1, 11111 and 1 1111159111111 40 0 11111 91nd 90 1131111151 g2. anu'v 1111 111 5 91' the 5 1111131121911 1115 h depe 11111911119. bedu- D111 0n J K police and. Am 5 C RPF BSF personnel depk Edna- yed in the Vane At the same D110 3 maimrsoxmelof Andhra ' police, mCl dmgGIlex Hound xx ere honoured Q g .; Ei E g reaunbnagnaipnvcuuutngmil

# Sphenis199ml my Obuse Window Help THRESH TRUNC 2

Fython 3.6.1 (v3.6.1:6900dbs, Mar 21 2017, 18:41:36) [Msc v.1900 64 bit (AMD64)] on win32

Type "copyright", "credits" or "license()" for more information. ======= RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ========= 70 Seconds Image to be processed: 2a.ipg h them 10 y 9 LL because 1111111111 1101 sure we would not need, 1' Educ 1 1110111 111111111 Pho es 11111111 111101119 e111cientand techm 111m; 1.1111111119311111191 1111.11911vmg110mewith- to ma, 1 011111 01111lng makes us anxious impor A batten 110 nmtter how sophisticated, is who 5 ew 11 11 When we have our cords with us we while 11.111 111 011113111111 pursuit. of power, even when pre 11 11V111111h'11111ugm. as a form of security. We 1f 011111 discover 0111 1111sfortune -- the loss of -1111191 m 1'1191111stoolate The opposite of how 1111111111111 harged 13 dead "Chet IV 111 011.1 lives ever be wire-free again?', 'podd Maybe not Apple introduced wireless ea ml phones last 1% and now Dell has a wireIGSs charging laptop but electric cars and their cimg rs am only beginning to arrive \_ For more The Verge 1 mage to be processed: 2b.jpg I'llig, UUI'SliVIS In part with them LIIIIIIIIII w 11111 not SIIH w WUHIH IIIII HILIKI 13.11 lll 1131111. Hunk IIIÎVI hIIUIUI IIIiiIIIIIII IIId 11112 IXIIiIII'iIIs I II 1111 )(i Io ninv III IIIIIWIIIL ' 11 I II II. ZIII' III ii'xis us. IIIXIIIIIS '.- II Iiitl IIIIIII INN iIOW sIIIIh 

Python 3.6.1 Shell THRESH\_TRUNC\_4 File Edit Shell Debug Options Window Help Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 Type "copyright", "credits" or "license()" for more information. [MSC v.1900 64 bit (AMD64)] on win32 ----- RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py -------19 seconds Image to be processed: 4a.jpg H S 3 I-Day Bareilly . I 9 I D I madrassas get govt warning Barely/Pilibhit The admi~ nistrat ion will take legal ac- tion against madrassas that dont, organise singing of na- tional anthem an record pro- ceedings on Independence Day. Bareilly divisional com- missionerPV-"Jagmohan said. We are Indians rst and our religion. caste or creed is secondary educational insti- tutes are cons idered public places. We will enforce q0 vemm em. order in madrassas and if them is blatant viola tion, we will take legal action, Jagmohan said. He added ac-v tion would bemkn only after going m Image to be processed: 4b.jpg 1-Day Bareilly 1tS madrassas get govt warning Bareilly/Pibhit The admi- nistmtion will take 1 egal 30 mm." against maclrassas that. dont cnganise singing Of na tirmal anthem and mxnd pm- (,irr xiings (m I nderxzendence 1.7); 3 Bmwilly divisional com-rwlissi ,;x1 rx P " .Jamnohzm said. We am Indizms rst. and our mligmn, caste or creed is ., . gxw n 1c121ry.. mlucxnkmal insti U " 1m m arr ( rimsiderml public ' ' plums. W will enfnrm g0~ -; wrmmm rnder m nwdrzgxssas i 8,1").d if ham is 1 ]laturit V1012 , tiem. we w i H rake hgzgz'sfi act ion Jagmohzm 5; id. 1 (73 mum 31 t; ion wuu Id be? w kf 51 only after Wink! 11 ];me Widmm Of Jiriitimw. TWPS 20% gallantry 3 s ; "i v w wumwm mother iii? medals to caps

## **Bibliography**

#### Patents:

- [1] US6577762B1, Background Surface thresholding
- [2] US7400768B1, Enhanced optical recognition of digitized images through selective bit-insertion.
- [3] US9298980B1, Image preprocessing for character recognition.
- [4] US20120063690A1, Object-Based Optical Character Recognition Pre-Processing Algorithm.
- [5] US7106905B2, Systems and methods for processing text-based electronic documents.
- [6] US20130329023A1, Text recognition driven functionality.

#### Literature:

- [7] Eugene Borovikov, A survey of modern optical character recognition techniques
- [8] M Seeger, C Dance, Binarising camera images for OCR (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)
- [9] Ranjith Unnikrishnan, Ray Smith, Combined Script and Page Orientation Estimation using the Tesseract OCR engine (ICDAR '07 Proceedings of the Ninth International Conference on Document Analysis and Recognition)
- [10] Ray Smith, An Overview of the Tesseract OCR Engine (MOCR '09 Proceedings of the International Workshop on Multilingual OCR)
- [11] Ray Smith, Daria Antoniva, Dar-Shyang Lee, *Adapting the Tesseract Open Source OCR Engine for Multilingual OCR* (MOCR '09 Proceedings of the International Workshop on Multilingual OCR)
- [12] Zheng Zhang, CL Tan, Binarizing document image using coplanar prefilter (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)
- [13] Zheng Zhang, CL Tan, Correcting document image warping based on regression of curved text lines (ICDAR 2003, Proceedings of the 9th International Conference on Document Analysis and Recognition)
- [14] Zheng Zhang, CL Tan, Recovery of distorted document images from bound volumes (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)