BVM Engineering College, VV Nagar

Gujarat Technological University

PROJECT #1

SEM 7
PRESENTATION

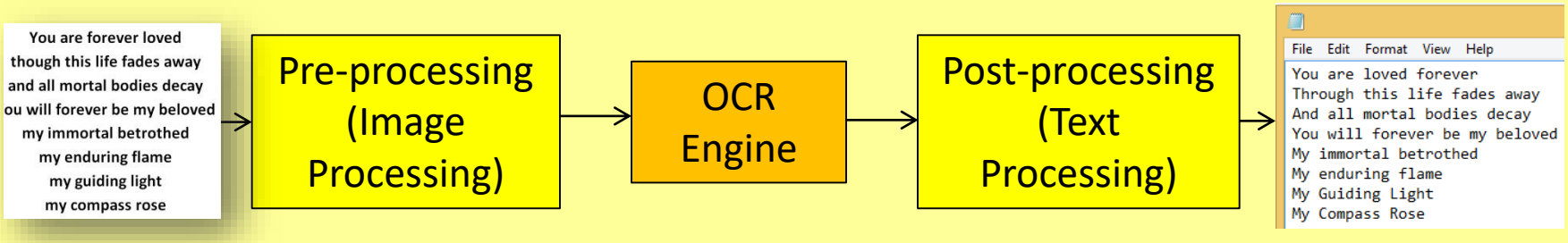
OCR-based Information Assistant

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.

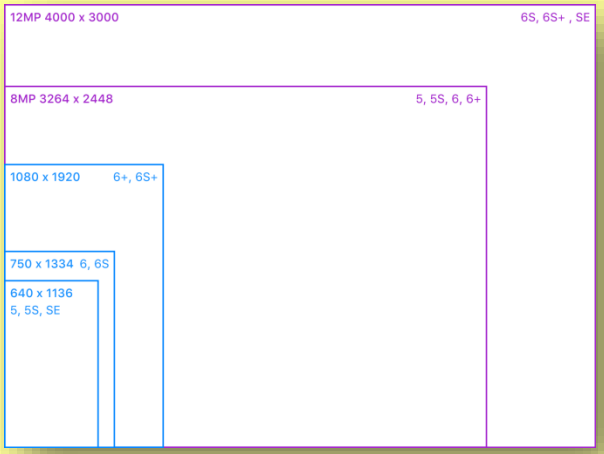
Steps Associated with Pre-Processing

Image processing

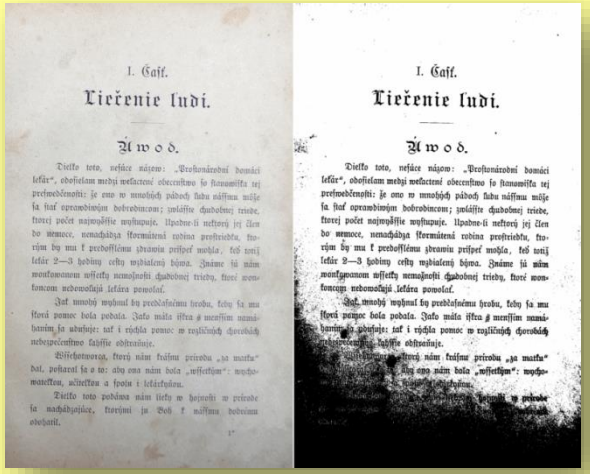
- Rescaling
- Binarisation (Thresholding)
- Rotation (De-Skewing)
- Border Removal
- Noise Removal

Page Segmentation

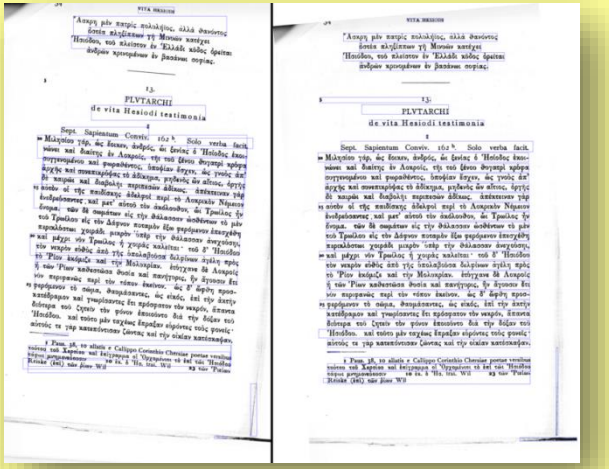
Dictionaries, Word lists & Patterns



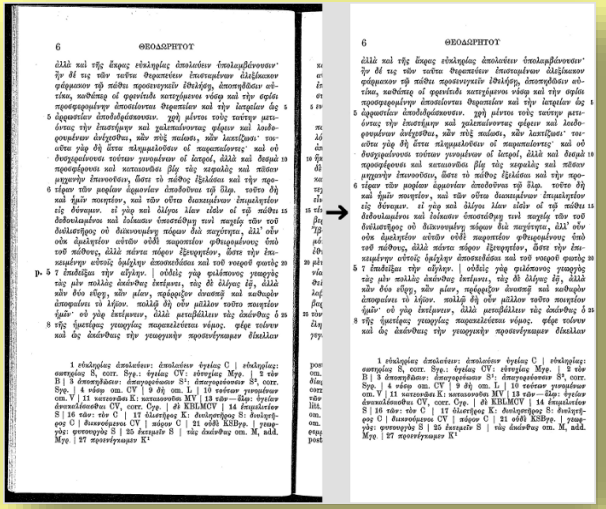
Rescaling



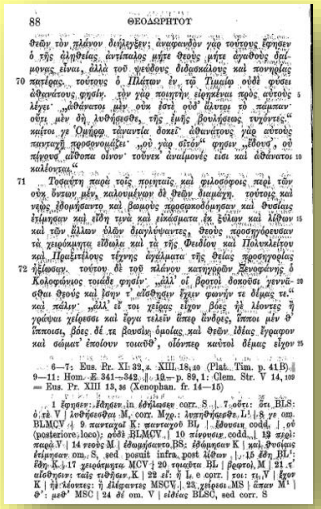
Binarisation



Rotation



Border Removal



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

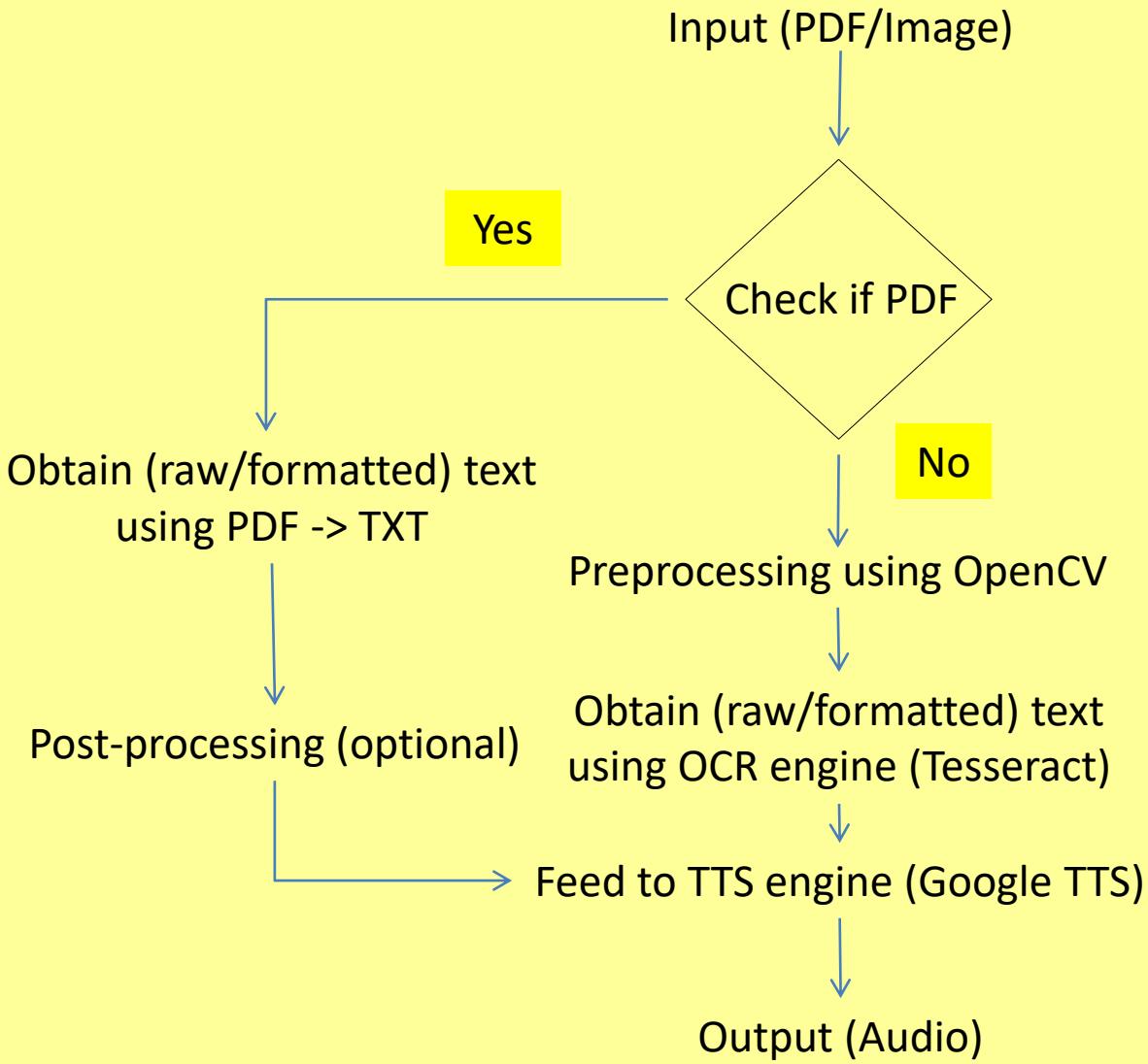
File Edit Format View Help

```
  You are foreVer 1oVeD
th0ugh 1his life fades away
and a11 morta1 bodies decay
ou will forever be m belove
 my immorta1 bet...
  my enduring f1ame
   my guid1ng 11ght
  my compas$ r0se
```

File Edit Format View Help

```
You are loved forever
Through this life fades away
And all mortal bodies decay
You will forever be my beloved
My immortal betrothed
My enduring flame
My Guiding Light
My Compass Rose
```

OCR-PIA: Flowchart

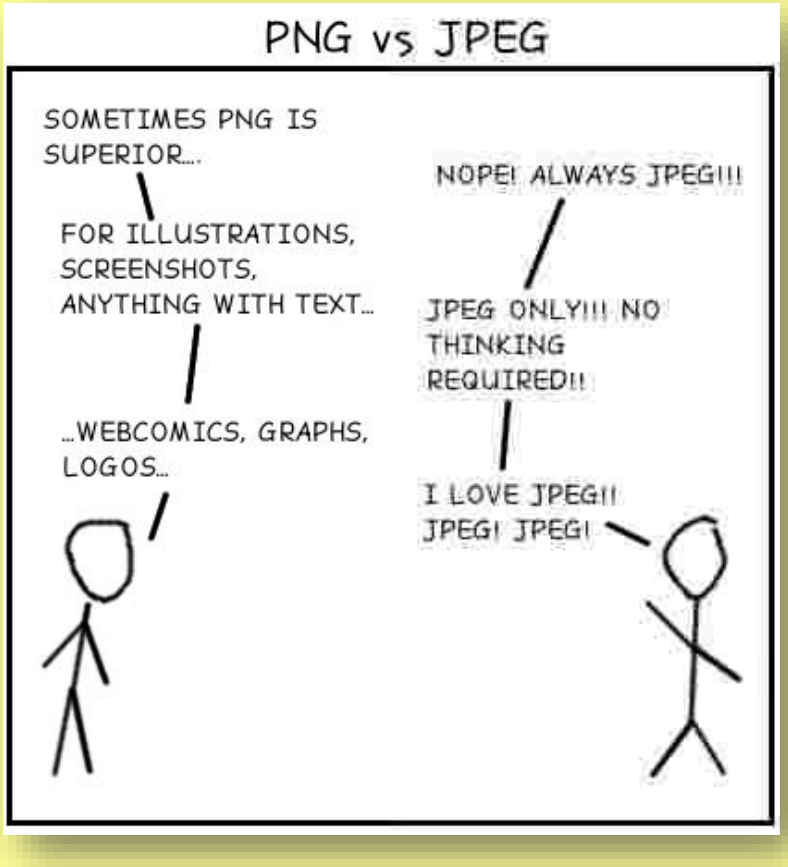


To JPEG or PNG, that is the question

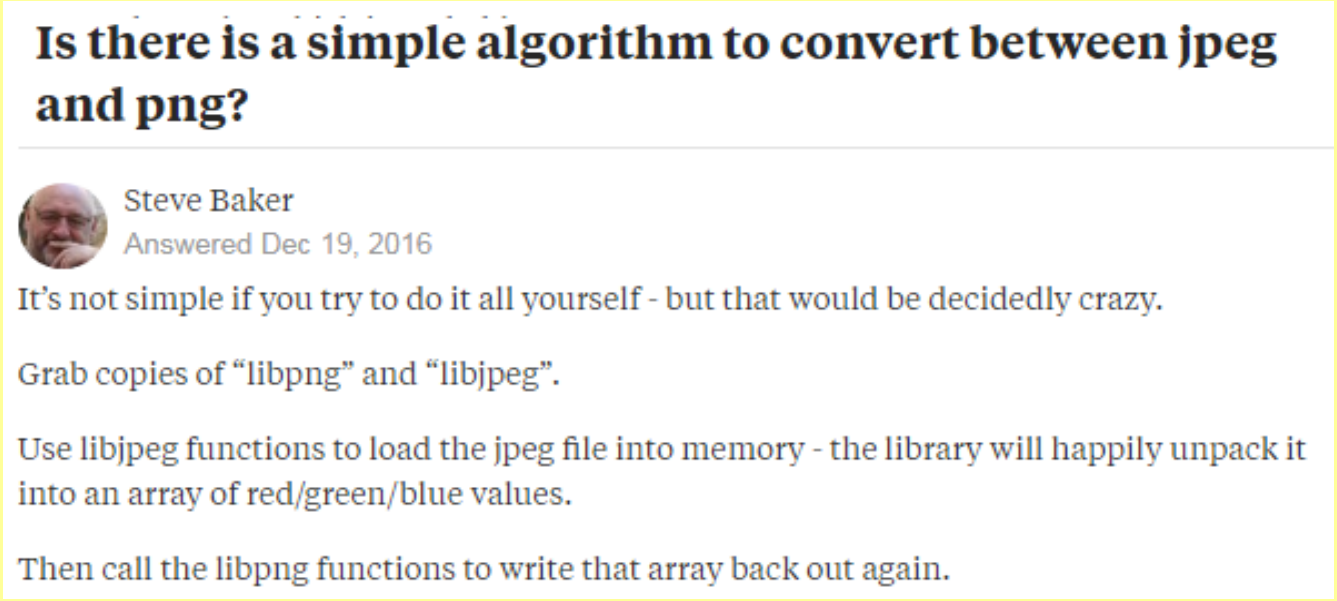
- JPEG uses lossy compression,PNG uses lossless compression.

WHAT DOES THIS MEAN?

- JPEG was designed for compressing full-colour/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

Technology has entangled our lives in wires

Have you ever thrown away a dead gadget, say an old tablet or an external hard drive, but kept its cables and charger? You are not alone.

"Just look in any household junk drawer and you'll find cords that have come and gone: relics of the camcorder era, laptops and digital cameras past, and outdated phone dongles," says an article in *The Verge*.

We can't bring ourselves to part with them because we are not sure we would not need them again. Phones have become efficient and their batteries larger, yet leaving home without a charger makes us anxious.

"A battery, no matter how sophisticated, is fleeting. When we have our cords with 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 it's too late. The opposite of being fully charged is dead."

Will our lives ever be wire-free again? Maybe not. Apple introduced wireless earphones last year, and now Dell has a wireless charging laptop, but electric cars and their chargers are only beginning to arrive.

Technology has entangled our lives in wires

Have you ever thrown away a dead gadget, say an old tablet or an external hard drive, but kept its cables and charger? You are not alone.

"Just look in any household junk drawer and you'll find cords that have come and gone: relics of the camcorder era, laptops and digital cameras past, and outdated phone dongles," says an article in *The Verge*.

We can't bring ourselves to part with them because we are not sure we would not need them again. Phones have become efficient and their batteries larger, yet leaving home without a charger makes us anxious.

"A battery, no matter how sophisticated, is fleeting. When we have our cords with 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 it's too late. The opposite of being fully charged is dead."

Will our lives ever be wire-free again? Maybe not. Apple introduced wireless earphones last year, and now Dell has a wireless charging laptop, but electric cars and their chargers are only beginning to arrive.

For more: The Verge

ed a generation

20% gallantry medals to cops for J&K service

The Centre, in appreciation of forces engaged in sustained counter-terror operations in J&K, on Monday honoured two CRPF commandants with Kirti Chakra for acts of bravery in J&K and conferred 40 of the 190 police gallantry medals awarded this Independence Day on J&K police and CRPF/BSF personnel deployed in the Valley. At the same time, 52 personnel of Andhra police, including Grey Hound commandos, were honoured with police gallantry medals for killing a record 24 Maoists in a single operation in October 2016. P8

I-Day: Bareilly madrassas get govt 'warning'

Bareilly/Pilibhit: The administration will "take legal action" against madrassas that don't organise singing of national anthem and record proceedings on Independence Day, Bareilly divisional commissioner P V Jagmohan said.

"We are Indians first and our religion, caste or creed is secondary... educational institutes are considered public places. We will enforce government order in madrassas and if there is blatant violation, we will take legal action," Jagmohan said. He added action would be taken only after going through "evidence of violation". P8

20% gallantry medals to cops

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

BINARY

BINARY_INV

TRUNC

TOZERO

TOZERO_INV

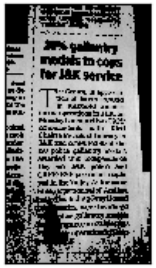
Original Image

BINARY

BINARY_INV

TRUNC

TOZERO



TOZERO_INV

Original Image

BINARY

BINARY_INV

TRUNC

TOZERO

TOZERO_INV

Original Image

BINARY

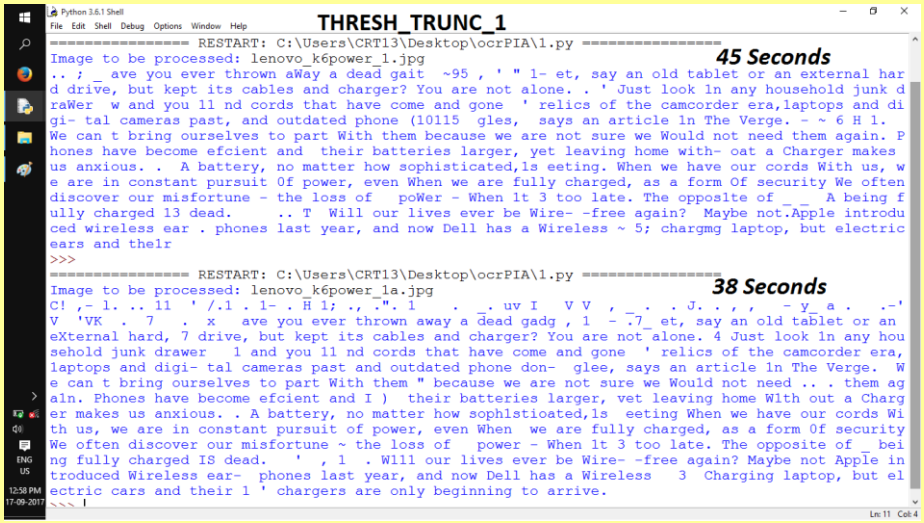
BINARY_INV

TRUNC

TOZERO

TOZERO_INV

THRESH_TRUNC_1

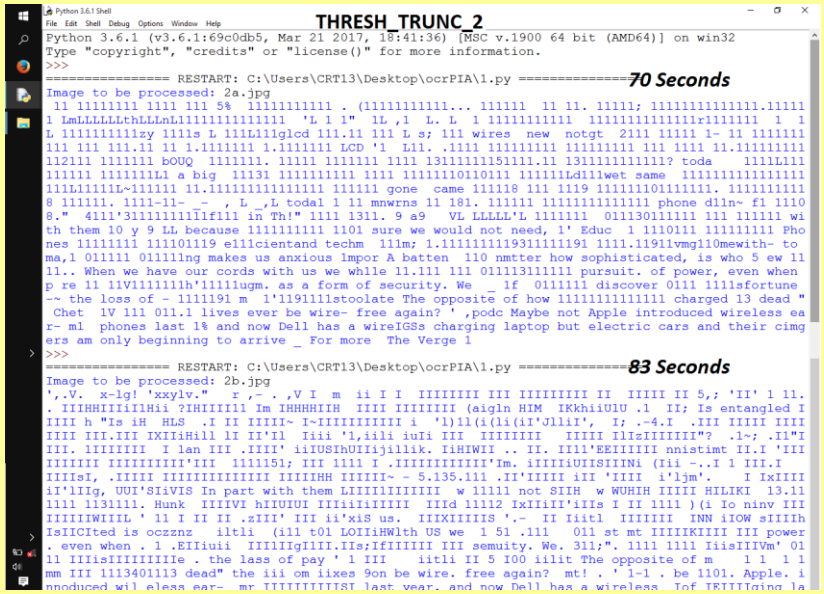


```
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: lenovo_k6power_1.jpg
.. ; _ ave you ever thrown aWay a dead gait  ~95 , ' " 1- et, say an old tablet or an external har
d drive, but kept its cables and charger? You are not alone. . ' Just look 1n any household 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 1n 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,ls 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 1t 3 too late. The opposlte 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 thelr
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: lenovo_k6power_1a.jpg
C! ,- l. .. 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 1n The Verge.  W
e 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, yet leaving home W1th out a Charg
er makes us anxious. . A battery, no matter how sophlstioated,ls  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 0f 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  . W111 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 el
ectric cars and their 1 ' chargers are only beginning to arrive.
>>>
```

45 Seconds

38 Seconds

THRESH_TRUNC_2

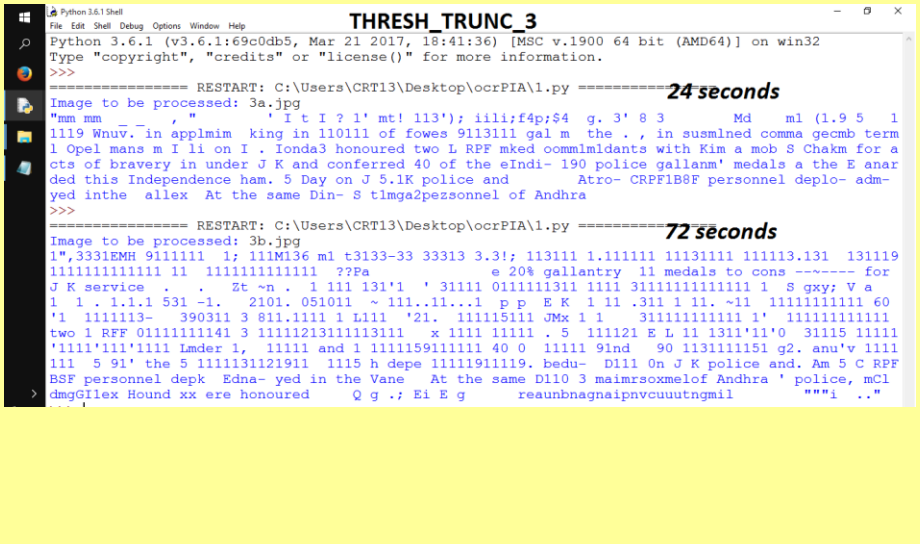


```
Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: 2a.jpg
 11 11111111 1111 111 5%  11111111111 . (11111111111... 111111  11 11. 11111; 111111111111111.11111
1 LmLLLLLLthLLLnL1111111111111111  'L 1 1"  1L ,1  L. L  1 1111111111  11111111111111r1111111  1  1
L 1111111111zy 1111s L 111L111glcd 111.11 111 L s; 111 wires  new  notgt  2111 11111 1- 11 1111111
111 111 111.11 11 1.1111111 1.1111111 LCD '1  L11. .1111 111111111 111111111 11 1111 11.111111111
112111 1111111 bOUQ  1111111. 11111 1111111 1111 13111111151111.11 1311111111111? toda   1111L111
111111 1111111L1 a big  11131 111111111 1111 11111110110111 111111Ld111wet same  11111111111111
111L11111L~111111 11.111111111111111 111111 gone  came 111118 111 1119 111111101111111. 1111111111
8 111111. 1111-11- _-  , L _,L todal 1 11 mnwrns 11 181. 111111 11111111111111 phone d11n~ f1 1110
8."  4111'3111111111f111 in Th!" 1111 1311. 9 a9  VL LLLLL'L 1111111  011130111111 111 111111 wi
th them 10 y 9 LL because 11111111111 1101 sure we would not need, 1' Educ  1 1110111 111111111 Pho
nes 11111111 111101119 e111cientand techm  111m; 1.11111111193111111191 1111.11911vmg110mewith- to
ma,1 011111 011111ng makes us anxious lmpor A batten  110 nmtter how sophisticated, is who 5 ew 11
11.. When we have our cords with us we whlle 11.111 111 011113111111 pursuit. of power, even when
p re 11 11V1111111h'11111ugm. as a form of security. We  _ 1f  0111111 discover 0111 1111sfortune
~~ the loss of - 1111191 m  1'1191111stoolate The opposite of how 11111111111111 charged 13 dead "
 Chet  1V 111 011.1 lives ever be wire- free again? ' ,podc Maybe not Apple introduced wireless ea
r- m1  phones last 1% and now Dell has a wireIGSs charging laptop but electric cars and their cimg
ers am only beginning to arrive _ For more  The Verge 1
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: 2b.jpg
',.V.  x-lg! 'xxylv."   r ,- . ,V I  m  ii I I  IIIIIIII III IIIIIIIII II  IIIII II 5,; 'II' 1 11.
. IIIHHIIIiIIHii ?IHIIIII1 Im IHHHHIIH  IIII IIIIIIII (aigln HIM  IKkhiiUlU .1  II; Is entangled I
IIII h "Is iH  HLS  .I II IIIII~ I~IIIIIIIIII i  'l)1l(i(li(iI'JlliI',  I; .-4.I  .III IIIII IIII
IIII III.III IXIIiHill lI II'Il  Iiii 'l,iili iuli III  IIIIIIII  IIIII IllzIIIIII"?  .l~; .II"I
III. IIIIIIII  I lan III .IIII' iiIUSIhUIIijillik. IiHIWII .. II. IIII'EEIIIII nnistimt II.I 'III
IIIIIII IIIIIIIIII'III  11111151; III 1111 I .IIIIIIIIIIII'Im. iIIIIiUIISIIINi (iii ~..I 1 III.I
IIIIsI, .IIIII IIIIIIIIIIIII IIIIHH IIIIII~ ~ 5.135.111 .II'IIIII iII 'IIII  i'ljm'.  I IxIIII
iI'IIIg, UUI'SIiVIS In part with them LIIIIIIIIIII  w 11111 not SIIH  w WUHIH IIIII HILIKI  13.11
1111 1131111. Hunk  IIIIVI hIIUUI IIIiIiIIIII  IIId 11112 IxIIiII'iIIs I II 1111 )(i Io ninv III
IIIIIIWIIIL ' 11 I II II .zIII' III ii'xi8 us.  IIIXIIIIS '.~  II Iiitl  IIIIIII  INN iIOW sIIIIh
IsIICIted is ocsznz  iltli  (111 t01 LOIIiHWlth U8 we  1 51 .111  011 st mt IIIIIKIIII III power
. even when . l .EIIiuiii  IIIiIIgI1II.IIs;IfIIIIII III semuity. We. 311;". 1111 1111 IiisIIIVm' 01
11 IIIisIIIIIIIIIe . the lass of pay ' 1 III  iitli II 5 I00 iilit The opposite of m  1 1  1 1
mm III 11134011113 dead" the iii om iixes 9on be wire. free again?  mt! . ' 1-1 . be 1101. Apple. i
nnoduced wil elass ear-  mr IIIIIIIIIIIIST last year, and now Dell has a wireless  Iof IKIIIIging la
```

70 Seconds

83 Seconds

THRESH_TRUNC_3

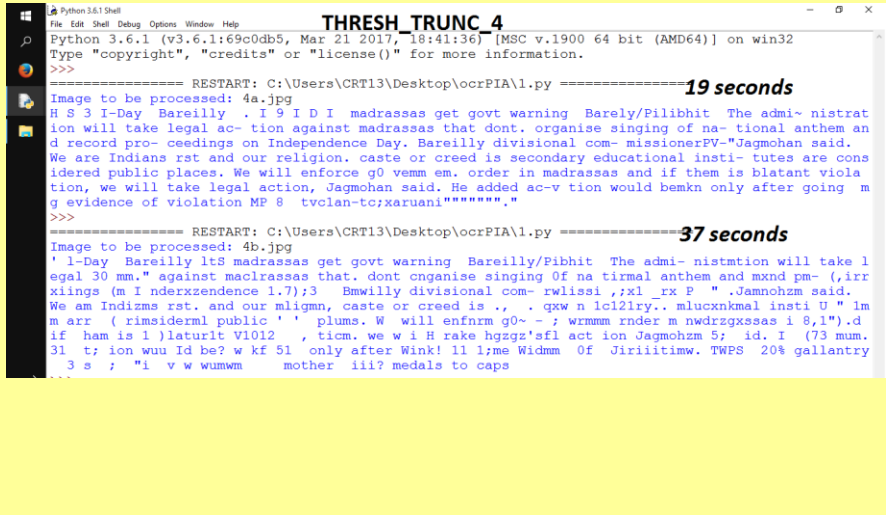


```
Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: 3a.jpg
"mm mm   _ _  , "       ' I t I ? 1' mt! 113'); iili;f4p;$4  g. 3' 8 3        Md    m1 (1.9 5   1
1119 Wnuv. in applmim  king in 110111 of fowes 9113111 gal m  the . , in susmlned comma gecmb term
l Opel mans m I li on I . Ionda3 honoured two L RPF mked oommlmldants 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- adm-
yed inthe  allex  At the same Din- S tlmga2pezsonnel of Andhra
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: 3b.jpg
1",3331EMH 9111111  1; 111M136 m1 t3133-33 33313 3.3!; 113111 1.111111 11131111 111113.131  131119
1111111111111 11  1111111111111  ??Pa        e 20% gallantry  11 medals to cons ~~~~~~ for
J K service  .  .   Zt ~n .  1 111 131'1  ' 31111 0111111311 1111 311111111111111 1  S gxy; V a
1  1 . 1.1.1 531 -1.   2101. 051011  ~ 111..11...1  p p  E K  1 11 .311 1 11. ~11  11111111111 60
'1  1111113-   390311 3 811.1111 1 L111  '21.  111115111 JMx 1 1    311111111111 1'  111111111111
two 1 RFF 011111111141 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
dmgG1ex Hound xx ere honoured     Q g .; Ei E g      reaunbnagnaipnvcuuutngmil       """i  .."
```

24 seconds

72 seconds

THRESH_TRUNC_4



```
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 ==========
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
d 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 g0 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
g evidence of violation MP 8  tvclan-tc;xaruani"""""""."
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: 4b.jpg
' l-Day  Bareilly ltS madrassas get govt warning  Bareilly/Pibhit  The admi- nistmtion will take l
egal 30 mm." against maclrassas that. dont cnganise singing 0f na tirmal anthem and mxnd pm- (,irr
xiings (m I nderxzendence 1.7);3  Bmwilly divisional com- rwlissi ,;xl _rx P  " .Jamnohzm said.
We am Indizms rst. and our mligmn, caste or creed is .,  . qxw n lcl21ry.. 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 )latur1t V1012  , ticm. we w i H rake hgzgz'sfl act ion Jagmohzm 5;  id. I  (73 mum.
31  t; ion wuu Id be? w kf 51  only after Wink! 11 1;me Widmm  0f  Jiriiitimw. TWPS  20% gallantry
  3 s  ;  "i  v w wumwm     mother  iii? medals to caps
```

19 seconds

37 seconds

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)