A
Project Report On

OCR-based Personal Assistant

Project-1 (2170001)

Bachelor of Engineering
in
Electronics & Communication Engineering

By

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

Under The Guidance of
Dr. Bhargav Goradiya
HoD, EC Department.

ELECTRONICS & COMMUNICATION ENGINEERING
DEPARTMENT
BVM ENGINEERING COLLEGE
GUJARAT TECHNOLOGICAL UNIVERSITY
VALLABH VIDYANAGAR-388120
Academic Year- 2017-18

INDEX

ACKNOWLEDGEMENT

We are extremely thankful to all our teachers, friends and fellow classmates for their continual support and assistance.

Chaitanya Tejaswi
Shahnawaz Yusufzai

CERTIFICATE

This is to certify that the project report entitled ***"OCR-based Personal Assistant",*** submitted by ***Chaitanya Tejaswi (140080111013) & Shahnawaz Yusufzai (120080112036)*** in the subject of the ***Project-1 (2170001)*** for the ***Bachelor of Engineering*** in ***Electronics & Communication*** of ***BVM Engineering College, Vallabh Vidyanagar (Gujarat Technological University)***, is the record of work carried out by them under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination.

Under The Guidance Of
Dr. Bhargav Goradiya
HoD, EC Department.

ELECTRONICS & COMMUNICATION ENGINEERING
DEPARTMENT
BVM ENGINEERING COLLEGE
GUJARAT TECHNOLOGICAL UNIVERSITY
VALLABH VIDYANAGAR-388120
Academic Year- 2017-18

PLAGIARISM REPORT

PlagScan Results of plagiarism analysis from 2017-09-30 12:15 UTC

Sem7-Project.docx

12.5%

Date: 2017-09-30 12:07 UTC

All sources **40** | Internet sources **9**

- [35] https://misteroleg.wordpress.com/2012/12...pre-processing-task/
  3.9% 12 matches
- [37] https://stackoverflow.com/questions/37378052/how-can-i-ocr-a-file-in-farsi
  3.8% 7 matches
- [39] https://www.quora.com/Where-can-I-get-th...e-with-documentation
  2.8% 7 matches
- [40] https://www.scribd.com/document/258117998/REport
  2.8% 4 matches
- [41] https://medium.com/@hdinhofer/optical-ch...-vision-76887e1d6ab0
  2.4% 3 matches
- [42] https://www.quora.com/How-does-the-Tesseract-API-for-OCR-work
  2.4% 2 matches
- [43] https://www.coursehero.com/file/pgpmjj/N...Tessesract-We-first/
  2.1% 4 matches
- [45] enlighten-ing.com/wp/
  0.4% 1 matches
- [46] ufology.wikia.com/wiki/Charles_Brown
  0.3% 1 matches

17 pages, 1803 words

A very light text-color was detected that might conceal letters used to merge words.

PlagLevel: selected / overall

133 matches from 47 sources, of which 45 are online sources.

Settings

Data policy: *Compare with web sources, Check against my documents*
Sensitivity: *Medium*
Bibliography: *Consider text*
Citation detection: *Reduce PlagLevel*
Whitelist: --

GUJARAT TECHNOLOGICAL UNIVERSITY
Undertaking: Originality of Work

We hereby certify that we are the sole authors of this UDP project report and that neither any part of this UDP project report nor the whole of the UDP Project report has been submitted for a degree by other student(s) to any other University or Institution.

We certify that, to the best of our knowledge, the current UDP Project report does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations or any other material from the work of other people included in our UDP Project report, published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

Furthermore, to the extent that we have included copyrighted material that surpasses the boundary of fair dealing within the meaning of the *Indian Copyright (Amendment) Act 2012*, we certify that we have obtained a written permission from the copyright owner(s) to include such material(s) in the current UDP Project report and have included copies of such copyright clearances to our appendix. We have checked the write up of the present UDP Project report using anti-plagiarism database and it is in the allowable limit.

In case of any complaints pertaining to plagiarism, we certify that we shall be solely responsible for the same and we understand that as per norms, University can even revoke BE degree conferred upon the student(s) submitting this IDP/UDP Project report, in case it is found to be plagiarized.

Enrollment No.	Name	Signature
140080111013	Chaitanya Tejaswi	
120080112036	Shahnawaz Yusufzai	

Dr. Bhargav Goradiya
(Project Guide)

Prof. Darshan Dalwadi
(Project Co-ordinator)

Introduction

What is a Personal Info Assistant (PIA)?
Following the idea of a "Personal Digital Assistant" (PDA), a PIA can be defined as a hardware which runs applications that provide quick reference to lists and processed data through proper links.

Why PIA?
We live in the age of information.
In the entire length of time between waking up to the sound of an alarm set on our branded smartphones and setting the same alarm before going to bed at night, we encounter a wide variety of tasks every day.
Common among these activities is the fact that each of these activities expects us to be informed. Using a washing machine needs us to know how to operate the buttons. Using an air-conditioner needs us to know what buttons to press on the remote in order to get the right setting.

Well, these are simple, aren't they?
Yes, because manufacturers make their products easy to use by hiding their inner features.

What's your response when your washing machine wouldn't run no matter how many buttons you press or your AC won't cool at the right temperature?
We all get irritated, don't we? The best response we have is to call *customer care* or a *repairman*.
Right at that moment something crosses my mind. Given our dependency on these, how little do we know about the things around us!
How much do we know about our 32" LCD Plasma TV; about the 5-speed automatic washing machine; about the 4-star rated AC; about the RO Water Purifier – all of which we operate on a daily basis at our homes?
"I'm not a techie guy", you may say, and you're correct. It's not necessary to knew everything about them, but a working knowledge wouldn't hurt.
This issue is serious for students & professionals.

Not getting the desired output on the CRO no matter how well you have connected the circuits? Does it bother you why it happened?
Been there. Done that.
The more we go out, the more we learn how little we know. But that is no excuse to have no idea about not having any idea about things we use on a daily basis.

The only problem is – **How would it be done?**

No one has the time or desire to leaf through the pages of a user guide; or to search for authentic documentation on the device.
Well, what if I told you that you do not need to do any of these; the information will reveal itself to you!
What would your response be?

Well, this project may do just that.

Tesseract-OCR

- Summary of Tesseract OCR Engine [9][10][11]
    1. Improving the quality of the output
    2. Tesseract-OCR : Control Parameters

Summary of Tesseract OCR Engine

Tesseract was originally developed at Hewlett-Packard Laboratories Bristol and at Hewlett-Packard Co, Greeley Colorado between 1985 and 1994, with some more changes made in 1996 to port to Windows, and some C++izing in 1998. In 2005 Tesseract was open sourced by HP. Since 2006 it is developed by Google.

- The 'tesseract-ocr'package contains an OCR engine - libtesseract and a command line program - tesseract.
- The lead developer is Ray Smith. The maintainer is Zdenko Podobny.
- Tesseract has unicode (UTF-8) support, and can recognize more than 100 languages "out of the box".
- Tesseract supports various output formats: plain-text, hocr(html), pdf, tsv, invisible-text-only pdf.
- You should note that in many cases, in order to get better OCR results, you'll need to improve the quality of the image you are giving Tesseract.
- This project does not include a GUI application. If you need one, please see the 3rdParty wiki page.
- Tesseract can be trained to recognize other languages. See Tesseract Training for more information.

Improving the quality of the output:

There are a variety of reasons you might not get good quality output from Tesseract. It's important to note that unless you're using a very unusual font or a new language retraining Tesseract is unlikely to help.

Image processing

Rescaling
Binarisation
Noise Removal
Rotation / Deskewing
Border Removal
Tools / Libraries

Page segmentation method

Dictionaries, word lists, and patterns

Image processing

Tesseract does various image processing operations internally (using the Leptonica library) before doing the actual OCR. It generally does a very good job of this, but there will inevitably be cases where it isn't good enough, which can result in a significant reduction in accuracy.

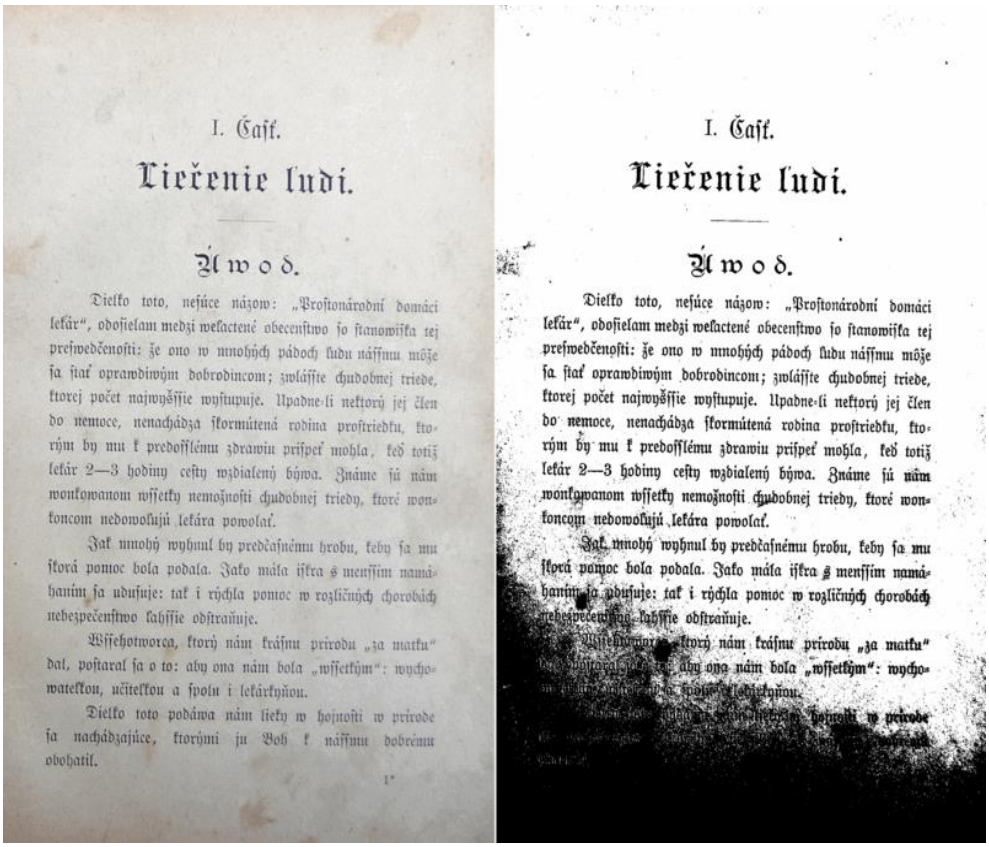
You can see how Tesseract has processed the image by using the configuration variable *tessedit_write_images* to true when running Tesseract. If the resulting tessinput.tif file looks problematic, try some of these image processing operations before passing the image to Tesseract.

Rescaling

Tesseract works best on images which have a DPI of at least 300 dpi, so it may be beneficial to resize images.

Binarisation

This is converting an image to black and white. Tesseract does this internally, but the result can be suboptimal, particularly if the page background is of uneven darkness.



I. Časť.

Liečenie ľudí.

Úwod.

Dielko toto, nesúce názow: „Prostonárodní domáci lekár", odosielam medzi welactené obecenstwo so stanowiska tej prešwedčenosti: že ono w mnohých pádoch ľudu náššmu môže sa stať oprawdiwým dobrodincom; zwlášťe chudobnej triede, ktorej počet najwyššie wystupuje. Upadne-li nektorý jej člen do nemoce, nenachádza skormútená rodina prostriedku, ktorým by mu k predošslému zdraviu prispeť mohla, keď totiž lekár 2—3 hodiny cesty wzdialený býwa. Známe sú nám wonkowanom wšetky nemožnosti chudobnej triedy, ktoré wonkoncom nedowolujú lekára powolať.

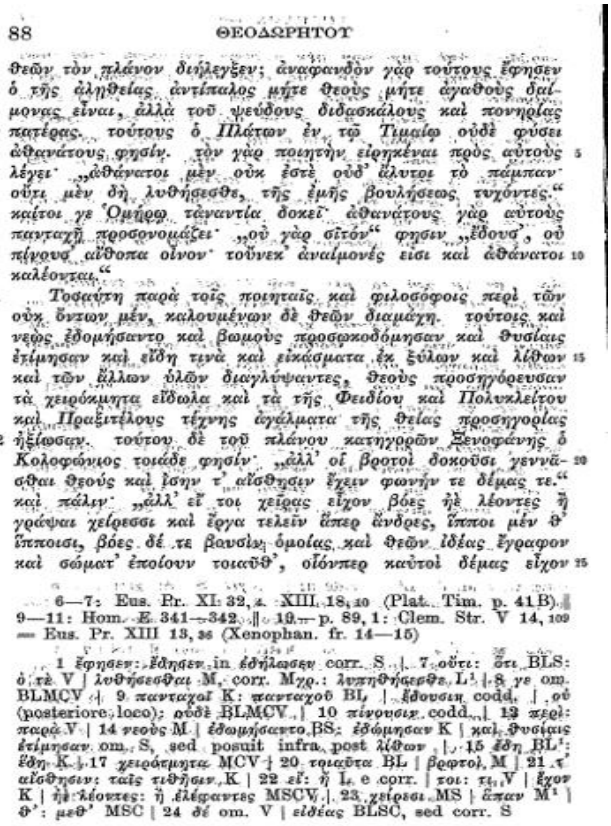
Jak mnohý wyhnul by predčasnému hrobu, keby sa mu skorá pomoc bola podala. Jako mála iskra s menšším namáhaním sa udušuje: tak i rýchla pomoc w rozličných chorobách nebezpečenstwo lahšie odstraňuje.

Wšehotworca, ktorý nám krásnu prirodu „za matku" dal, postaral sa o to: aby ona nám bola „wšetkým": wychowateľkou, učiteľkou a spolu i lekárkyňou.

Dielko toto podáwa nám lieky w hojnosti w prirode sa nachádzajúce, ktorými ju Boh k náššmu dobrému obohatil.

Noise Removal

Noise is random variation of brightness or colour in an image, that can make the text of the image more difficult to read. Certain types of noise cannot be removed by Tesseract in the binarisation step, which can cause accuracy rates to drop.



88 ΘΕΟΔΩΡΗΤΟΥ

θεῶν τὸν πλάνον διήλεγξεν; ἀναφανδὸν γὰρ τούτους ἔφησεν ὁ τῆς ἀληθείας ἀντίπαλος μήτε θεοὺς μήτε ἀγαθοὺς δαίμονας εἶναι, ἀλλὰ τοῦ ψεύδους διδασκάλους καὶ πονηρίας πατέρας. τούτους ὁ Πλάτων ἐν τῷ Τιμαίῳ οὐδὲ φύσει ἀθανάτους φησίν. τὸν γὰρ ποιητὴν εἰρηκέναι πρὸς αὐτοὺς λέγει· „ἀθάνατοι μὲν οὐκ ἐστὲ οὐδ' ἄλυτοι τὸ παμπαν· οὔτι μὲν δὴ λυθήσεσθε, τῆς ἐμῆς βουλήσεως τυχόντες." καίτοι γε Ὁμήρῳ τἀναντία δοκεῖ· ἀθανάτους γὰρ αὐτοὺς πανταχῇ προσονομάζει· „οὐ γὰρ σῖτον" φησιν „ἔδουσ', οὐ πίνουσ' αἴθοπα οἶνον· τοὔνεκ' ἀναίμονές εἰσι καὶ ἀθάνατοι καλέονται."

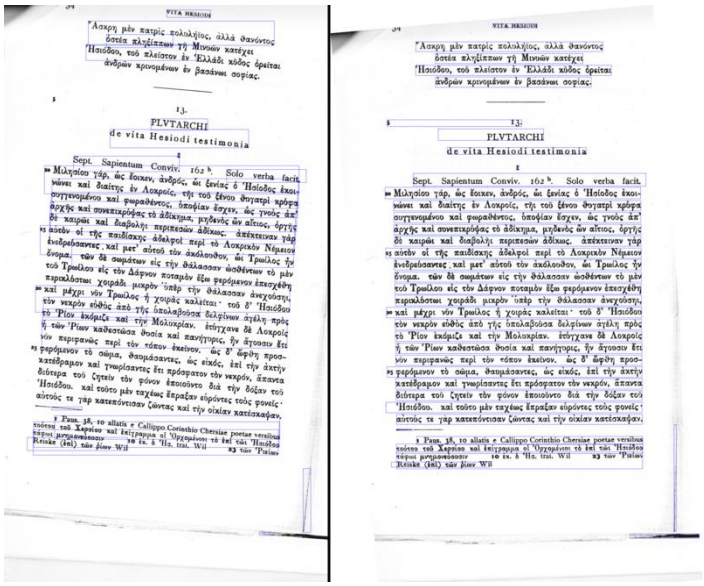
Τοσαύτη παρὰ τοῖς ποιηταῖς καὶ φιλοσόφοις περὶ τῶν οὐκ ὄντων μέν, καλουμένων δὲ θεῶν διαμάχη. τούτοις καὶ νεῴς ἐδομήσαντο καὶ βωμοὺς προσῳκοδόμησαν καὶ θυσίαις ἐτίμησαν καὶ εἴδη τινὰ καὶ εἰκάσματα ἐκ ξύλων καὶ λίθων καὶ τῶν ἄλλων ὑλῶν διαγλύψαντες, θεοὺς προσηγόρευσαν τὰ χειρόκμητα εἴδωλα καὶ τὰ τῆς Φειδίου καὶ Πολυκλείτου καὶ Πραξιτέλους τέχνης ἀγάλματα τῆς θείας προσηγορίας ἠξίωσαν. τούτου δὲ τοῦ πλάνου κατηγορῶν Ξενοφάνης ὁ Κολοφώνιος τοιάδε φησίν· „ἀλλ' οἱ βροτοὶ δοκοῦσι γεννᾶσθαι θεούς καὶ ἴσην τ' αἴσθησιν ἔχειν φωνήν τε δέμας τε." καὶ πάλιν· „ἀλλ' εἴ τοι χεῖρας εἶχον βόες ἠὲ λέοντες ἢ γράψαι χείρεσσι καὶ ἔργα τελεῖν ἅπερ ἄνδρες, ἵπποι μέν θ' ἵπποισι, βόες δέ τε βουσὶν ὁμοίας καὶ θεῶν ἰδέας ἔγραφον καὶ σώματ' ἐποίουν τοιαῦθ', οἷόνπερ καὐτοὶ δέμας εἶχον

6—7: Eus. Pr. XI 32, 4. XIII 18, 10 (Plat. Tim. p. 41 B). 9—11: Hom. E 341—342. 12—p. 89, 1: Clem. Str. V 14, 109 = Eus. Pr. XIII 13, 36 (Xenophan. fr. 14—15)

1 ἔφησεν: ἔδησεν in ἐσήλωσεν corr. S | 7 οὔτι: ὅτι BLS: ὁ τὲ V | λυθήσεσθαι M, corr. Mχρ.: λυπηθήσεσθε L¹ | 8 γε om. BLMCV | 9 πανταχοῖ K: πανταχοῦ BL | ἔδουσιν codd. | οὐ (posteriore loco): οὐδὲ BLMCV | 10 πίνουσιν codd. | 12 περὶ: παρὰ V | 14 νεοὺς M | ἐδωμήσαντο BS: ἐδώμησαν K | καὶ θυσίαις ἐτίμησαν om. S, sed posuit infra post λίθων | 15 εἴδη BL¹: ἔδη K | 17 χειρόκμητα MCV | 20 τοιαῦτα BL | βροτοὶ M | 21 τ' αἴσθησιν: ταῖς τιθήσιν K | 22 εἶ: ἢ L, e corr. | τοι: τι V | ἔχον K | ἠὲ λέοντες: ἢ ἐλέφαντες MSCV | 23 χείρεσι MS | ἅπαν M¹ | θ': μεθ' MSC | 24 δέ om. V | εἰδέας BLSC, sed corr. S

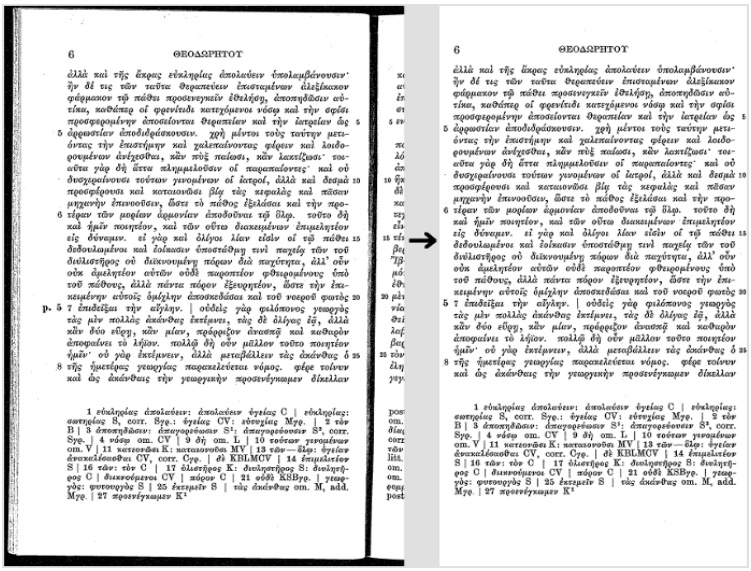
Rotation / Deskewing

A skewed image is when an page has been scanned when not straight. The quality of Tesseract's line segmentation reduces significantly if a page is too skewed, which severely impacts the quality of the OCR. To address this rotating the page image so that the text lines are horizontal.



Border Removal

Scanned pages often have dark borders around them. These can be erroneously picked up as extra characters, especially if they vary in shape and gradation.



Tools / Libraries

- Leptonica
- OpenCV
- Scan Tailor
- ImageMagick
- unpaper
- ImageJ
- Gimp

Page segmentation method

By default Tesseract expects a page of text when it segments an image. If you're just seeking to OCR a small region try a different segmentation mode, using the -psm argument. Note that adding a white border to text which is too tightly cropped may also help, see issue 398.
To see a complete list of supported page segmentation modes, use tesseract -h. Here's the list as of 3.21:

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.

11 Sparse text. Find as much text as possible in no particular order.

12 Sparse text with OSD.

13 Raw line. Treat the image as a single text line, bypassing hacks that are Tesseract-specific.

Dictionaries, word lists, and patterns

By default Tesseract is optimized to recognize sentences of words. If you're trying to recognize something else, like receipts, price lists, or codes, there are a few things you can do to improve the accuracy of your results, as well as double-checking that the appropriate segmentation method is selected.

Disabling the dictionaries Tesseract uses should increase recognition if most of your text isn't dictionary words. They can be disabled by setting the both of the configuration variables load_system_dawg and load_freq_dawg to false.

It is also possible to add words to the word list Tesseract uses to help recognition, or to add common character patterns, which can further help to improve accuracy if you have a good idea of the sort of input you expect. This is explained in more detail in the Tesseract manual.

If you know you will only encounter a subset of the characters available in the language, such as only digits, you can use the tessedit_char_whitelist configuration variable. See the FAQ for an example.

Tesseract: Control Parameters

```
C:\Users\CRT13>tesseract --print-parameters

Tesseract parameters:
classify_num_cp_levels  3       Number of Class Pruner Levels
textord_dotmatrix_gap   3       Max pixel gap for broken pixed pitch
textord_debug_block     0       Block to do debug on
textord_pitch_range     2       Max range test on pitch
textord_words_veto_power          5         Rows required to outvote a
veto
textord_tabfind_show_strokewidths       0       Show stroke widths
pitsync_linear_version  6       Use new fast algorithm
pitsync_fake_depth      1       Max advance fake generation
oldbl_holed_losscount   10      Max lost before fallback line used
textord_skewsmooth_offset       4       For smooth factor
textord_skewsmooth_offset2      1       For smooth factor
textord_test_x  -2147483647     coord of test pt
textord_test_y  -2147483647     coord of test pt
textord_min_blobs_in_row           4         Min blobs before gradient
counted
textord_spline_minblobs 8       Min blobs in each spline segment
textord_spline_medianwin           6         Size of window for spline
segmentation
textord_max_blob_overlaps       4       Max number of blobs a big blob
can overlap
textord_min_xheight     10      Min credible pixel xheight
textord_lms_line_trials 12      Number of linew fits to do
textord_tabfind_show_images     0       Show image blobs
textord_fp_chop_error   2       Max allowed bending of chop cells
edges_max_children_per_outline  10       Max number of children inside
a character outline
edges_max_children_layers        5       Max layers of nested children
inside a character outline
edges_children_per_grandchild   10       Importance ratio for chucking
outlines
edges_children_count_limit      45      Max holes allowed in blob
edges_min_nonhole       12      Min pixels for potential char in box
edges_patharea_ratio     40       Max lensq/area for acceptable child
outline
devanagari_split_debuglevel      0        Debug level for split shiro-
rekha process.
textord_tabfind_show_partitions 0       Show partition bounds, waiting
if >1
textord_debug_tabfind   0       Debug tab finding
textord_debug_bugs       0       Turn on output related to bugs in tab
finding
textord_testregion_left -1      Left edge of debug reporting rectangle
textord_testregion_top  -1      Top edge of debug reporting rectangle
```

```
textord_testregion_right          2147483647       Right edge of debug
rectangle
textord_testregion_bottom         2147483647       Bottom edge of debug
rectangle
editor_image_xpos       590     Editor image X Pos
editor_image_ypos       10      Editor image Y Pos
editor_image_menuheight 50      Add to image height for menu bar
editor_image_word_bb_color      7       Word bounding box colour
editor_image_blob_bb_color      4       Blob bounding box colour
editor_image_text_color 2       Correct text colour
editor_dbwin_xpos       50      Editor debug window X Pos
editor_dbwin_ypos       500     Editor debug window Y Pos
editor_dbwin_height     24      Editor debug window height
editor_dbwin_width      80      Editor debug window width
editor_word_xpos        60      Word window X Pos
editor_word_ypos        510     Word window Y Pos
editor_word_height      240     Word window height
editor_word_width       655     Word window width
wordrec_display_splits  0       Display splits
poly_debug      0       Debug old poly
poly_wide_objects_better          1        More accurate approx on wide
things
wordrec_display_all_blobs       0       Display Blobs
wordrec_display_all_words       0       Display Words
wordrec_blob_pause      0       Blob pause
textord_fp_chopping     1       Do fixed pitch chopping
textord_force_make_prop_words     0           Force proportional word
segmentation on all rows
textord_chopper_test    0       Chopper is being tested.
textord_restore_underlines      1       Chop underlines & put back
textord_show_initial_words      0       Display separate words
textord_show_new_words  0       Display separate words
textord_show_fixed_words          0         Display forced fixed pitch
words
textord_blocksall_fixed 0       Moan about prop blocks
textord_blocksall_prop  0       Moan about fixed pitch blocks
textord_blocksall_testing       0       Dump stats when moaning
textord_test_mode       0       Do current test
textord_pitch_scalebigwords     0       Scale scores on big words
textord_all_prop        0       All doc is proportial text
textord_debug_pitch_test        0       Debug on fixed pitch test
textord_disable_pitch_test         0          Turn off dp fixed pitch
algorithm
textord_fast_pitch_test 0       Do even faster pitch algorithm
textord_debug_pitch_metric      0       Write full metric stuff
textord_show_row_cuts   0       Draw row-level cuts
textord_show_page_cuts  0       Draw page-level cuts
textord_pitch_cheat     0       Use correct answer for fixed/prop
textord_blockndoc_fixed 0       Attempt whole doc/block fixed pitch
textord_dump_table_images       0       Paint table detection output
```

```
textord_show_tables     0       Show table regions
textord_tablefind_show_mark      0        Debug table marking steps in
detail
textord_tablefind_show_stats     0        Show page stats used in table
finding
textord_tablefind_recognize_tables        0          Enables the table
recognizer for table layout and filtering.
textord_tabfind_show_initialtabs        0       Show tab candidates
textord_tabfind_show_finaltabs  0       Show tab vectors
textord_tabfind_only_strokewidths       0       Only run stroke widths
textord_really_old_xheight      0       Use original wiseowl xheight
textord_oldbl_debug     0       Debug old baseline generation
textord_debug_baselines 0       Debug baseline generation
textord_oldbl_paradef   1       Use para default mechanism
textord_oldbl_split_splines     1       Split stepped splines
textord_oldbl_merge_parts       1       Merge suspect partitions
oldbl_corrfix   1       Improve correlation of heights
oldbl_xhfix     0       Fix bug in modes threshold for xheights
textord_ocropus_mode    0       Make baselines for ocropus
textord_heavy_nr        0       Vigorously remove noise
textord_show_initial_rows       0       Display row accumulation
textord_show_parallel_rows      0       Display page correlated rows
textord_show_expanded_rows      0       Display rows after expanding
textord_show_final_rows 0       Display rows after final fitting
textord_show_final_blobs         0       Display blob bounds after pre-
ass
textord_test_landscape  0       Tests refer to land/port
textord_parallel_baselines      1       Force parallel baselines
textord_straight_baselines      0       Force straight baselines
textord_old_baselines   1       Use old baseline algorithm
textord_old_xheight     0       Use old xheight algorithm
textord_fix_xheight_bug 1       Use spline baseline
textord_fix_makerow_bug 1       Prevent multiple baselines
textord_debug_xheights  0       Test xheight algorithms
textord_biased_skewcalc 1       Bias skew estimates with line length
textord_interpolating_skew      1       Interpolate across gaps
textord_new_initial_xheight     1       Use test xheight mechanism
textord_debug_blob      0       Print test blob information
gapmap_debug    0       Say which blocks have tables
gapmap_use_ends 0       Use large space at start and end of rows
gapmap_no_isolated_quanta         0         Ensure gaps not less than
2quanta wide
edges_use_new_outline_complexity          0        Use the new outline
complexity module
edges_debug     0       turn on debugging for this module
edges_children_fix        0          Remove boxy parents of char-like
children
textord_show_fixed_cuts 0       Draw fixed pitch cell boundaries
devanagari_split_debugimage       0         Whether to create a debug
image for split shiro-rekha process.
```

```
textord_tabfind_show_color_fit  0       Show stroke widths
textord_tabfind_show_initial_partitions 0       Show partition bounds
textord_tabfind_show_reject_blobs       0       Show blobs rejected as
noise
textord_tabfind_show_columns    0       Show column bounds
textord_tabfind_show_blocks     0       Show final block bounds
textord_tabfind_find_tables     1       run table detection
textord_space_size_is_variable  0       If true, word delimiter spaces
are assumed to have variable width, even though characters have fixed
pitch.
textord_debug_images    0       Use greyed image background for debug
textord_debug_printable 0       Make debug windows printable
equationdetect_save_bi_image    0       Save input bi image
equationdetect_save_spt_image   0       Save special character image
equationdetect_save_seed_image  0       Save the seed image
equationdetect_save_merged_image        0       Save the merged image
stream_filelist 0       Stream a filelist from stdin
debug_file              File to send tprintf output to
classify_training_file  MicroFeatures   Training file
classify_font_name       UnknownFont      Default font name to be used
in training
fx_debugfile    FXDebug Name of debugfile
editor_image_win_name   EditorImage     Editor image window name
editor_dbwin_name       EditorDBWin     Editor debug window name
editor_word_name        BlnWords        BL normalized word window
editor_debug_config_file                Config file to apply to single
words
classify_pico_feature_length    0.05    Pico Feature Length
classify_norm_adj_midpoint      32      Norm adjust midpoint ...
classify_norm_adj_curl  2       Norm adjust curl ...
classify_min_slope       0.414214         Slope below which lines are
called horizontal
classify_max_slope        2.41421 Slope above which lines are called
vertical
classify_cp_angle_pad_loose     45      Class Pruner Angle Pad Loose
classify_cp_angle_pad_medium    20      Class Pruner Angle Pad Medium
classify_cp_angle_pad_tight     10      CLass Pruner Angle Pad Tight
classify_cp_end_pad_loose       0.5     Class Pruner End Pad Loose
classify_cp_end_pad_medium      0.5     Class Pruner End Pad Medium
classify_cp_end_pad_tight       0.5     Class Pruner End Pad Tight
classify_cp_side_pad_loose      2.5     Class Pruner Side Pad Loose
classify_cp_side_pad_medium     1.2     Class Pruner Side Pad Medium
classify_cp_side_pad_tight      0.6     Class Pruner Side Pad Tight
classify_pp_angle_pad   45      Proto Pruner Angle Pad
classify_pp_end_pad     0.5     Proto Prune End Pad
classify_pp_side_pad    2.5     Proto Pruner Side Pad
textord_underline_offset        0.1     Fraction of x to ignore
textord_wordstats_smooth_factor 0.05    Smoothing gap stats
textord_width_smooth_factor     0.1     Smoothing width stats
textord_words_width_ile 0.4     Ile of blob widths for space est
```

```
textord_words_maxspace  4       Multiple of xheight
textord_words_default_maxspace  3.5     Max believable third space
textord_words_default_minspace  0.6     Fraction of xheight
textord_words_min_minspace      0.3     Fraction of xheight
textord_words_default_nonspace  0.2     Fraction of xheight
textord_words_initial_lower     0.25    Max initial cluster size
textord_words_initial_upper     0.15    Min initial cluster spacing
textord_words_minlarge  0.75    Fraction of valid gaps needed
textord_words_pitchsd_threshold 0.04    Pitch sync threshold
textord_words_def_fixed 0.016   Threshold for definite fixed
textord_words_def_prop  0.09    Threshold for definite prop
textord_pitch_rowsimilarity        0.08      Fraction  of  xheight  for
sameness
words_initial_lower     0.5     Max initial cluster size
words_initial_upper     0.15    Min initial cluster spacing
words_default_prop_nonspace     0.25    Fraction of xheight
words_default_fixed_space       0.75    Fraction of xheight
words_default_fixed_limit       0.6     Allowed size variance
textord_words_definite_spread   0.3     Non-fuzzy spacing region
textord_spacesize_ratiofp       2.8     Min ratio space/nonspace
textord_spacesize_ratioprop     2       Min ratio space/nonspace
textord_fpiqr_ratio     1.5     Pitch IQR/Gap IQR threshold
textord_max_pitch_iqr   0.2     Xh fraction noise in pitch
textord_fp_min_width    0.5     Min width of decent blobs
textord_projection_scale        0.2     Ding rate for mid-cuts
textord_balance_factor  1       Ding rate for unbalanced char cells
textord_tabvector_vertical_gap_fraction 0.5      max fraction of mean
blob width allowed for vertical gaps in vertical text
textord_tabvector_vertical_box_ratio       0.5        Fraction  of  box
matches required to declare a line vertical
pitsync_joined_edge     0.75    Dist inside big blob for chopping
pitsync_offset_freecut_fraction 0.25    Fraction of cut for free cuts
oldbl_xhfract   0.4     Fraction of est allowed in calc
oldbl_dot_error_size    1.26    Max aspect ratio of a dot
textord_oldbl_jumplimit 0.15    X fraction for new partition
textord_spline_shift_fraction    0.02     Fraction of line spacing for
quad
textord_spline_outlier_fraction 0.1      Fraction of line spacing for
outlier
textord_skew_ile        0.5     Ile of gradients for page skew
textord_skew_lag        0.02    Lag for skew on row accumulation
textord_linespace_iqrlimit      0.2     Max iqr/median for linespace
textord_width_limit     8       Max width of blobs to make rows
textord_chop_width      1.5     Max width before chopping
textord_expansion_factor          1        Factor to expand rows by in
expand_rows
textord_overlap_x       0.375   Fraction of linespace for good overlap
textord_minxh   0.25    fraction of linesize for min xheight
textord_min_linesize    1.25    * blob height for initial linesize
```

```
textord_excess_blobsize 1.3      New row made if blob makes row this
big
textord_occupancy_threshold     0.4     Fraction of neighbourhood
textord_underline_width 2       Multiple of line_size for underline
textord_min_blob_height_fraction        0.75    Min blob height/top to
include blob top into xheight stats
textord_xheight_mode_fraction     0.4        Min  pile  height  to  make
xheight
textord_ascheight_mode_fraction  0.08       Min  pile  height  to  make
ascheight
textord_descheight_mode_fraction          0.08     Min pile height to
make descheight
textord_ascx_ratio_min  1.25    Min cap/xheight
textord_ascx_ratio_max  1.8     Max cap/xheight
textord_descx_ratio_min 0.25    Min desc/xheight
textord_descx_ratio_max 0.6     Max desc/xheight
textord_xheight_error_margin    0.1     Accepted variation
gapmap_big_gaps 1.75    xht multiplier
textord_fp_chop_snap    0.5     Max distance of chop pt from vertex
edges_childarea 0.5     Min area fraction of child outline
edges_boxarea   0.875   Min area fraction of grandchild for box
textord_underline_threshold     0.5     Fraction of width occupied
ambigs_debug_level      0       Debug level for unichar ambiguities
tessedit_single_match   0       Top choice only from CP
classify_debug_level    0       Classify debug level
classify_norm_method    1       Normalization Method   ...
matcher_debug_level     0       Matcher Debug Level
matcher_debug_flags     0       Matcher Debug Flags
classify_learning_debug_level   0       Learning Debug Level:
matcher_permanent_classes_min   1       Min # of permanent classes
matcher_min_examples_for_prototyping      3           Reliable  Config
Threshold
matcher_sufficient_examples_for_prototyping        5            Enable
adaption even if the ambiguities have not been seen
classify_adapt_proto_threshold   230        Threshold  for  good  protos
during adaptive 0-255
classify_adapt_feature_threshold          230      Threshold for good
features during adaptive 0-255
classify_class_pruner_threshold 229     Class Pruner Threshold 0-255
classify_class_pruner_multiplier             15          Class  Pruner
Multiplier 0-255:
classify_cp_cutoff_strength     7       Class Pruner CutoffStrength:
classify_integer_matcher_multiplier        10          Integer  Matcher
Multiplier  0-255:
il1_adaption_test         0         Don't adapt to i/I at beginning of
word
dawg_debug_level         0        Set to 1 for general debug info, to 2
for more details, to 3 to see all the debug messages
hyphen_debug_level      0       Debug level for hyphenated words.
max_viterbi_list_size   10      Maximum size of viterbi list.
```

stopper_smallword_size 2 Size of dict word to be treated as non-dict word
stopper_debug_level 0 Stopper debug level
tessedit_truncate_wordchoice_log 10 Max words to keep in list
fragments_debug 0 Debug character fragments
max_permuter_attempts 10000 Maximum number of different character choices to consider during permutation. This limit is especially useful when user patterns are specified, since overly generic patterns can result in dawg search exploring an overly large number of options.
repair_unchopped_blobs 1 Fix blobs that aren't chopped
chop_debug 0 Chop debug
chop_split_length 10000 Split Length
chop_same_distance 2 Same distance
chop_min_outline_points 6 Min Number of Points on Outline
chop_seam_pile_size 150 Max number of seams in seam_pile
chop_inside_angle -50 Min Inside Angle Bend
chop_min_outline_area 2000 Min Outline Area
chop_centered_maxwidth 90 Width of (smaller) chopped blobs above which we don't care that a chop is not near the center.
chop_x_y_weight 3 X / Y length weight
segment_adjust_debug 0 Segmentation adjustment debug
wordrec_debug_level 0 Debug level for wordrec
wordrec_max_join_chunks 4 Max number of broken pieces to associate
segsearch_debug_level 0 SegSearch debug level
segsearch_max_pain_points 2000 Maximum number of pain points stored in the queue
segsearch_max_futile_classifications 20 Maximum number of pain point classifications per chunk thatdid not result in finding a better word choice.
language_model_debug_level 0 Language model debug level
language_model_ngram_order 8 Maximum order of the character ngram model
language_model_viterbi_list_max_num_prunable 10 Maximum number of prunable (those for which PrunablePath() is true) entries in each viterbi list recorded in BLOB_CHOICEs
language_model_viterbi_list_max_size 500 Maximum size of viterbi lists recorded in BLOB_CHOICEs
language_model_min_compound_length 3 Minimum length of compound words
wordrec_display_segmentations 0 Display Segmentations
tessedit_pageseg_mode 6 Page seg mode: 0=osd only, 1=auto+osd, 2=auto, 3=col, 4=block, 5=line, 6=word, 7=char (Values from PageSegMode enum in publictypes.h)
tessedit_ocr_engine_mode 0 Which OCR engine(s) to run (Tesseract, Cube, both). Defaults to loading and running only Tesseract (no Cube,no combiner). Values from OcrEngineMode enum in tesseractclass.h)

```
pageseg_devanagari_split_strategy         0         Whether to use the
top-line splitting process for Devanagari documents while performing
page-segmentation.
ocr_devanagari_split_strategy    0         Whether to use the top-line
splitting process for Devanagari documents while performing ocr.
bidi_debug      0       Debug level for BiDi
applybox_debug  1       Debug level
applybox_page   0       Page number to apply boxes from
tessedit_bigram_debug    0          Amount of debug output for bigram
correction.
debug_noise_removal     0       Debug reassignment of small outlines
noise_maxperblob        8       Max diacritics to apply to a blob
noise_maxperword        16      Max diacritics to apply to a word
debug_x_ht_level        0       Reestimate debug
quality_min_initial_alphas_reqd 2       alphas in a good word
tessedit_tess_adaption_mode      39       Adaptation decision algorithm
for tess
tessedit_test_adaption_mode      3        Adaptation decision algorithm
for tess
paragraph_debug_level   0       Print paragraph debug info.
cube_debug_level        0       Print cube debug info.
tessedit_preserve_min_wd_len     2        Only preserve wds longer than
this
crunch_rating_max       10      For adj length in rating per ch
crunch_pot_indicators   1       How many potential indicators needed
crunch_leave_lc_strings  4         Don't crunch words with long lower
case strings
crunch_leave_uc_strings  4         Don't crunch words with long lower
case strings
crunch_long_repetitions 3       Crunch words with long repetitions
crunch_debug    0       As it says
fixsp_non_noise_limit   1       How many non-noise blbs either side?
fixsp_done_mode 1       What constitues done for spacing
debug_fix_space_level   0       Contextual fixspace debug
x_ht_acceptance_tolerance        8        Max allowed deviation of blob
top outside of font data
x_ht_min_change 8       Min change in xht before actually trying it
superscript_debug         0         Debug level for sub & superscript
fixer
suspect_level   99      Suspect marker level
suspect_space_level     100     Min suspect level for rejecting spaces
suspect_short_words      2         Don't suspect dict wds longer than
this
tessedit_reject_mode    0       Rejection algorithm
tessedit_image_border   2       Rej blbs near image edge limit
min_sane_x_ht_pixels    8       Reject any x-ht lt or eq than this
tessedit_page_number    -1      -1 -> All pages , else specifc page to
process
tessdata_manager_debug_level         0               Debug  level  for
TessdataManager functions.
```

```
tessedit_parallelize    0       Run in parallel where possible
tessedit_ok_mode        5       Acceptance decision algorithm
segment_debug   0       Debug the whole segmentation process
language_model_fixed_length_choices_depth        3        Depth of blob
choice lists to explore when fixed length dawgs are on
tosp_debug_level        0       Debug data
tosp_enough_space_samples_for_median    3       or should we use mean
tosp_redo_kern_limit    10      No.samples reqd to reestimate for row
tosp_few_samples        40      No.gaps reqd with 1 large gap to treat
as a table
tosp_short_row  20      No.gaps reqd with few cert spaces to use certs
tosp_sanity_method      1       How to avoid being silly
textord_max_noise_size  7       Pixel size of noise
textord_baseline_debug  0       Baseline debug level
textord_noise_sizefraction      10      Fraction of size for maxima
textord_noise_translimit        16      Transitions for normal blob
textord_noise_sncount   1       super norm blobs to save row
use_definite_ambigs_for_classifier          0            Use  definite
ambiguities when running character classifier
use_ambigs_for_adaption 0         Use ambigs for deciding whether to
adapt to a character
allow_blob_division     1       Use divisible blobs chopping
prioritize_division     0       Prioritize blob division over chopping
classify_enable_learning        1       Enable adaptive classifier
tess_cn_matching        0       Character Normalized Matching
tess_bn_matching        0       Baseline Normalized Matching
classify_enable_adaptive_matcher            1          Enable  adaptive
classifier
classify_use_pre_adapted_templates         0          Use  pre-adapted
classifier templates
classify_save_adapted_templates 0         Save adapted templates to a
file
classify_enable_adaptive_debugger       0       Enable match debugger
classify_nonlinear_norm   0                 Non-linear   stroke-density
normalization
disable_character_fragments       1          Do not include character
fragments in the results of the classifier
classify_debug_character_fragments        0         Bring up graphical
debugging windows for fragments training
matcher_debug_separate_windows  0        Use two different windows for
debugging the matching: One for the protos and one for the features.
classify_bln_numeric_mode        0        Assume the input is numbers
[0-9].
load_system_dawg        1       Load system word dawg.
load_freq_dawg  1       Load frequent word dawg.
load_unambig_dawg       1       Load unambiguous word dawg.
load_punc_dawg  1       Load dawg with punctuation patterns.
load_number_dawg        1       Load dawg with number patterns.
load_bigram_dawg        1       Load dawg with special word bigrams.
```

use_only_first_uft8_step 0 Use only the first UTF8 step of the given string when computing log probabilities.
stopper_no_acceptable_choices 0 Make AcceptableChoice() always return false. Useful when there is a need to explore all segmentations
save_raw_choices 0 Deprecated- backward compatibility only
segment_nonalphabetic_script 0 Don't use any alphabetic-specific tricks.Set to true in the traineddata config file for scripts that are cursive or inherently fixed-pitch
save_doc_words 0 Save Document Words
merge_fragments_in_matrix 1 Merge the fragments in the ratings matrix and delete them after merging
wordrec_no_block 0 Don't output block information
wordrec_enable_assoc 1 Associator Enable
force_word_assoc 0 force associator to run regardless of what enable_assoc is.This is used for CJK where component grouping is necessary.
fragments_guide_chopper 0 Use information from fragments to guide chopping process
chop_enable 1 Chop enable
chop_vertical_creep 0 Vertical creep
chop_new_seam_pile 1 Use new seam_pile
assume_fixed_pitch_char_segment 0 include fixed-pitch heuristics in char segmentation
wordrec_skip_no_truth_words 0 Only run OCR for words that had truth recorded in BlamerBundle
wordrec_debug_blamer 0 Print blamer debug messages
wordrec_run_blamer 0 Try to set the blame for errors
save_alt_choices 1 Save alternative paths found during chopping and segmentation search
language_model_ngram_on 0 Turn on/off the use of character ngram model
language_model_ngram_use_only_first_uft8_step 0 Use only the first UTF8 step of the given string when computing log probabilities.
language_model_ngram_space_delimited_language 1 Words are delimited by space
language_model_use_sigmoidal_certainty 0 Use sigmoidal score for certainty
tessedit_resegment_from_boxes 0 Take segmentation and labeling from box file
tessedit_resegment_from_line_boxes 0 Conversion of word/line box file to char box file
tessedit_train_from_boxes 0 Generate training data from boxed chars
tessedit_make_boxes_from_boxes 0 Generate more boxes from boxed chars
tessedit_dump_pageseg_images 0 Dump intermediate images made during page segmentation
tessedit_ambigs_training 0 Perform training for ambiguities

tessedit_adaption_debug 0 Generate and print debug information for adaption
applybox_learn_chars_and_char_frags_mode 0 Learn both character fragments (as is done in the special low exposure mode) as well as unfragmented characters.
applybox_learn_ngrams_mode 0 Each bounding box is assumed to contain ngrams. Only learn the ngrams whose outlines overlap horizontally.
tessedit_display_outwords 0 Draw output words
tessedit_dump_choices 0 Dump char choices
tessedit_timing_debug 0 Print timing stats
tessedit_fix_fuzzy_spaces 1 Try to improve fuzzy spaces
tessedit_unrej_any_wd 0 Don't bother with word plausibility
tessedit_fix_hyphens 1 Crunch double hyphens?
tessedit_redo_xheight 1 Check/Correct x-height
tessedit_enable_doc_dict 1 Add words to the document dictionary
tessedit_debug_fonts 0 Output font info per char
tessedit_debug_block_rejection 0 Block and Row stats
tessedit_enable_bigram_correction 1 Enable correction based on the word bigram dictionary.
tessedit_enable_dict_correction 0 Enable single word correction based on the dictionary.
enable_noise_removal 1 Remove and conditionally reassign small outlines when they confuse layout analysis, determining diacritics vs noise
debug_acceptable_wds 0 Dump word pass/fail chk
tessedit_minimal_rej_pass1 0 Do minimal rejection on pass 1 output
tessedit_test_adaption 0 Test adaption criteria
tessedit_matcher_log 0 Log matcher activity
test_pt 0 Test for point
paragraph_text_based 1 Run paragraph detection on the post-text-recognition (more accurate)
docqual_excuse_outline_errs 0 Allow outline errs in unrejection?
tessedit_good_quality_unrej 1 Reduce rejection on good docs
tessedit_use_reject_spaces 1 Reject spaces?
tessedit_preserve_blk_rej_perfect_wds 1 Only rej partially rejected words in block rejection
tessedit_preserve_row_rej_perfect_wds 1 Only rej partially rejected words in row rejection
tessedit_dont_blkrej_good_wds 0 Use word segmentation quality metric
tessedit_dont_rowrej_good_wds 0 Use word segmentation quality metric
tessedit_row_rej_good_docs 1 Apply row rejection to good docs
tessedit_reject_bad_qual_wds 1 Reject all bad quality wds
tessedit_debug_doc_rejection 0 Page stats

```
tessedit_debug_quality_metrics  0       Output data to debug file
bland_unrej     0       unrej potential with no chekcs
unlv_tilde_crunching    1       Mark v.bad words for tilde crunch
hocr_font_info  0       Add font info to hocr output
crunch_early_merge_tess_fails   1       Before word crunch?
crunch_early_convert_bad_unlv_chs       0       Take out ~^ early?
crunch_terrible_garbage 1       As it says
crunch_pot_garbage      1       POTENTIAL crunch garbage
crunch_leave_ok_strings 1       Don't touch sensible strings
crunch_accept_ok        1       Use acceptability in okstring
crunch_leave_accept_strings       0          Don't pot crunch sensible
strings
crunch_include_numerals 0       Fiddle alpha figures
tessedit_prefer_joined_punct    0       Reward punctation joins
tessedit_write_block_separators  0           Write block separators in
output
tessedit_write_rep_codes        0       Write repetition char code
tessedit_write_unlv     0       Write .unlv output file
tessedit_create_txt     0       Write .txt output file
tessedit_create_hocr    0       Write .html hOCR output file
tessedit_create_tsv     0       Write .tsv output file
tessedit_create_pdf     0       Write .pdf output file
textonly_pdf    0       Create PDF with only one invisible text layer
suspect_constrain_1Il   0       UNLV keep 1Il chars rejected
tessedit_minimal_rejection      0       Only reject tess failures
tessedit_zero_rejection 0       Don't reject ANYTHING
tessedit_word_for_word  0        Make output have exactly one word per
WERD
tessedit_zero_kelvin_rejection  0       Don't reject ANYTHING AT ALL
tessedit_consistent_reps        1       Force all rep chars the same
tessedit_rejection_debug        0       Adaption debug
tessedit_flip_0O        1       Contextual 0O O0 flips
rej_trust_doc_dawg      0       Use DOC dawg in 11l conf. detector
rej_1Il_use_dict_word   0       Use dictword test
rej_1Il_trust_permuter_type     1       Don't double check
rej_use_tess_accepted   1       Individual rejection control
rej_use_tess_blanks     1       Individual rejection control
rej_use_good_perm       1       Individual rejection control
rej_use_sensible_wd     0       Extend permuter check
rej_alphas_in_number_perm       0       Extend permuter check
tessedit_create_boxfile 0       Output text with boxes
tessedit_write_images   0       Capture the image from the IPE
interactive_display_mode        0       Run interactively?
tessedit_override_permuter      1       According to dict_word
tessedit_use_primary_params_model        0        In multilingual mode
use params model of the primary language
textord_tabfind_show_vlines     0       Debug line finding
textord_use_cjk_fp_model        0       Use CJK fixed pitch model
poly_allow_detailed_fx   0         Allow feature extractors to see the
original outline
```

```
tessedit_init_config_only          0          Only initialize with the
config file. Useful if the instance is not going to be used for OCR
but say only for layout analysis.
textord_equation_detect 0       Turn on equation detector
textord_tabfind_vertical_text   1       Enable vertical detection
textord_tabfind_force_vertical_text      0        Force using vertical
text page mode
preserve_interword_spaces         0         Preserve multiple interword
spaces
include_page_breaks       0          Include page separator string in
output text after each image/page.
textord_tabfind_vertical_horizontal_mix 1        find horizontal lines
such as headers in vertical page mode
load_fixed_length_dawgs 1       Load fixed length dawgs (e.g. for non-
space delimited languages)
permute_debug   0       Debug char permutation process
permute_script_word       0          Turn on word script consistency
permuter
segment_segcost_rating  0        incorporate segmentation cost in word
rating?
permute_fixed_length_dawg            0            Turn on fixed-length
phrasebook search permuter
permute_chartype_word    0          Turn on character type (property)
consistency permuter
ngram_permuter_activated          0         Activate character-level n-
gram-based permuter
permute_only_top        0       Run only the top choice permuter
use_new_state_cost        0         use new state cost heuristics for
segmentation state evaluation
enable_new_segsearch    0       Enable new segmentation search path.
textord_single_height_mode       0        Script has no xheight, so use
a single mode
tosp_old_to_method      0       Space stats use prechopping?
tosp_old_to_constrain_sp_kn      0        Constrain relative values of
inter and intra-word gaps for old_to_method.
tosp_only_use_prop_rows 1       Block stats to use fixed pitch rows?
tosp_force_wordbreak_on_punct    0        Force word breaks on punct to
break long lines in non-space delimited langs
tosp_use_pre_chopping   0       Space stats use prechopping?
tosp_old_to_bug_fix     0       Fix suspected bug in old code
tosp_block_use_cert_spaces      1       Only stat OBVIOUS spaces
tosp_row_use_cert_spaces        1       Only stat OBVIOUS spaces
tosp_narrow_blobs_not_cert      1       Only stat OBVIOUS spaces
tosp_row_use_cert_spaces1       1       Only stat OBVIOUS spaces
tosp_recovery_isolated_row_stats          1         Use row alone when
inadequate cert spaces
tosp_only_small_gaps_for_kern   0       Better guess
tosp_all_flips_fuzzy    0       Pass ANY flip to context?
tosp_fuzzy_limit_all     1        Don't restrict kn->sp fuzzy limit to
tables
```

```
tosp_stats_use_xht_gaps 1       Use within xht gap for wd breaks
tosp_use_xht_gaps       1       Use within xht gap for wd breaks
tosp_only_use_xht_gaps  0       Only use within xht gap for wd breaks
tosp_rule_9_test_punct  0       Don't chng kn to space next to punct
tosp_flip_fuzz_kn_to_sp 1       Default flip
tosp_flip_fuzz_sp_to_kn 1       Default flip
tosp_improve_thresh     0       Enable improvement heuristic
textord_no_rejects      0       Don't remove noise blobs
textord_show_blobs      0       Display unsorted blobs
textord_show_boxes      0       Display unsorted blobs
textord_noise_rejwords  1       Reject noise-like words
textord_noise_rejrows   1       Reject noise-like rows
textord_noise_debug     0       Debug row garbage detector
m_data_sub_dir  tessdata/       Directory for data files
tessedit_module_name    libtesseract    Module colocated with tessdata
dir
classify_learn_debug_str                Class str to debug learning
user_words_file         A filename of user-provided words.
user_words_suffix                     A suffix of user-provided words
located in tessdata.
user_patterns_file              A filename of user-provided patterns.
user_patterns_suffix                A suffix of user-provided patterns
located in tessdata.
output_ambig_words_file           Output file for ambiguities found in
the dictionary
word_to_debug                Word for which stopper debug information
should be printed to stdout
word_to_debug_lengths           Lengths of unichars in word_to_debug
tessedit_char_blacklist         Blacklist of chars not to recognize
tessedit_char_whitelist         Whitelist of chars to recognize
tessedit_char_unblacklist                    List of chars to override
tessedit_char_blacklist
tessedit_write_params_to_file              Write all parameters to the
given file.
applybox_exposure_pattern         .exp     Exposure value follows this
pattern in the image filename. The name of the image files are
expected to be in the form [lang].[fontname].exp[num].tif
chs_leading_punct       ('`"    Leading punctuation
chs_trailing_punct1     ).,;:?! 1st Trailing punctuation
chs_trailing_punct2     )'`"    2nd Trailing punctuation
outlines_odd    %|      Non standard number of outlines
outlines_2      ij!?%":;        Non standard number of outlines
numeric_punctuation     .,      Punct. chs expected WITHIN numbers
unrecognised_char       |       Output char for unidentified blobs
ok_repeated_ch_non_alphanum_wds -?*=    Allow NN to unrej
conflict_set_I_l_1      Il1[]   Il1 conflict set
file_type       .tif    Filename extension
tessedit_load_sublangs            List of languages to load with this
one
page_separator
```

Page separator (default is form feed control character)
classify_char_norm_range 0.2 Character Normalization Range ...
classify_min_norm_scale_x 0 Min char x-norm scale ...
classify_max_norm_scale_x 0.325 Max char x-norm scale ...
classify_min_norm_scale_y 0 Min char y-norm scale ...
classify_max_norm_scale_y 0.325 Max char y-norm scale ...
classify_max_rating_ratio 1.5 Veto ratio between classifier ratings
classify_max_certainty_margin 5.5 Veto difference between classifier certainties
matcher_good_threshold 0.125 Good Match (0-1)
matcher_reliable_adaptive_result 0 Great Match (0-1)
matcher_perfect_threshold 0.02 Perfect Match (0-1)
matcher_bad_match_pad 0.15 Bad Match Pad (0-1)
matcher_rating_margin 0.1 New template margin (0-1)
matcher_avg_noise_size 12 Avg. noise blob length
matcher_clustering_max_angle_delta 0.015 Maximum angle delta for prototype clustering
classify_misfit_junk_penalty 0 Penalty to apply when a non-alnum is vertically out of its expected textline position
rating_scale 1.5 Rating scaling factor
certainty_scale 20 Certainty scaling factor
tessedit_class_miss_scale 0.00390625 Scale factor for features not used
classify_adapted_pruning_factor 2.5 Prune poor adapted results this much worse than best result
classify_adapted_pruning_threshold -1 Threshold at which classify_adapted_pruning_factor starts
classify_character_fragments_garbage_certainty_threshold -3 Exclude fragments that do not look like whole characters from training and adaption
speckle_large_max_size 0.3 Max large speckle size
speckle_rating_penalty 10 Penalty to add to worst rating for noise
xheight_penalty_subscripts 0.125 Score penalty (0.1 = 10%) added if there are subscripts or superscripts in a word, but it is otherwise OK.
xheight_penalty_inconsistent 0.25 Score penalty (0.1 = 10%) added if an xheight is inconsistent.
segment_penalty_dict_frequent_word 1 Score multiplier for word matches which have good case andare frequent in the given language (lower is better).
segment_penalty_dict_case_ok 1.1 Score multiplier for word matches that have good case (lower is better).
segment_penalty_dict_case_bad 1.3125 Default score multiplier for word matches, which may have case issues (lower is better).
segment_penalty_ngram_best_choice 1.24 Multipler to for the best choice from the ngram model.

segment_penalty_dict_nonword 1.25 Score multiplier for glyph fragment segmentations which do not match a dictionary word (lower is better).
segment_penalty_garbage 1.5 Score multiplier for poorly cased strings that are not in the dictionary and generally look like garbage (lower is better).
certainty_scale 20 Certainty scaling factor
stopper_nondict_certainty_base -2.5 Certainty threshold for non-dict words
stopper_phase2_certainty_rejection_offset 1 Reject certainty offset
stopper_certainty_per_char -0.5 Certainty to add for each dict char above small word size.
stopper_allowable_character_badness 3 Max certaintly variation allowed in a word (in sigma)
doc_dict_pending_threshold 0 Worst certainty for using pending dictionary
doc_dict_certainty_threshold -2.25 Worst certainty for words that can be inserted into thedocument dictionary
wordrec_worst_state 1 Worst segmentation state
tessedit_certainty_threshold -2.25 Good blob limit
chop_split_dist_knob 0.5 Split length adjustment
chop_overlap_knob 0.9 Split overlap adjustment
chop_center_knob 0.15 Split center adjustment
chop_sharpness_knob 0.06 Split sharpness adjustment
chop_width_change_knob 5 Width change adjustment
chop_ok_split 100 OK split limit
chop_good_split 50 Good split limit
segsearch_max_char_wh_ratio 2 Maximum character width-to-height ratio
language_model_ngram_small_prob 1e-006 To avoid overly small denominators use this as the floor of the probability returned by the ngram model.
language_model_ngram_nonmatch_score -40 Average classifier score of a non-matching unichar.
language_model_ngram_scale_factor 0.03 Strength of the character ngram model relative to the character classifier
language_model_ngram_rating_factor 16 Factor to bring log-probs into the same range as ratings when multiplied by outline length
language_model_penalty_non_freq_dict_word 0.1 Penalty for words not in the frequent word dictionary
language_model_penalty_non_dict_word 0.15 Penalty for non-dictionary words
language_model_penalty_punc 0.2 Penalty for inconsistent punctuation
language_model_penalty_case 0.1 Penalty for inconsistent case
language_model_penalty_script 0.5 Penalty for inconsistent script
language_model_penalty_chartype 0.3 Penalty for inconsistent character type

```
language_model_penalty_font     0       Penalty for inconsistent font
language_model_penalty_spacing    0.05       Penalty  for  inconsistent
spacing
language_model_penalty_increment        0.01    Penalty increment
noise_cert_basechar     -8      Hingepoint for base char certainty
noise_cert_disjoint     -1      Hingepoint for disjoint certainty
noise_cert_punc -3      Threshold for new punc char certainty
noise_cert_factor           0.375     Scaling  on  certainty  diff  from
Hingepoint
quality_rej_pc  0.08    good_quality_doc lte rejection limit
quality_blob_pc 0       good_quality_doc gte good blobs limit
quality_outline_pc        1         good_quality_doc  lte  outline  error
limit
quality_char_pc 0.95    good_quality_doc gte good char limit
test_pt_x       100000  xcoord
test_pt_y       100000  ycoord
tessedit_reject_doc_percent      65       %rej allowed before rej whole
doc
tessedit_reject_block_percent    45       %rej allowed before rej whole
block
tessedit_reject_row_percent      40       %rej allowed before rej whole
row
tessedit_whole_wd_rej_row_percent        70       Number of row rejects
in whole word rejectswhich prevents whole row rejection
tessedit_good_doc_still_rowrej_wd         1.1       rej  good  doc  wd  if
more than this fraction rejected
quality_rowrej_pc       1.1     good_quality_doc gte good char limit
crunch_terrible_rating  80      crunch rating lt this
crunch_poor_garbage_cert        -9      crunch garbage cert lt this
crunch_poor_garbage_rate        60      crunch garbage rating lt this
crunch_pot_poor_rate    40      POTENTIAL crunch rating lt this
crunch_pot_poor_cert    -8      POTENTIAL crunch cert lt this
crunch_del_rating       60      POTENTIAL crunch rating lt this
crunch_del_cert -10     POTENTIAL crunch cert lt this
crunch_del_min_ht       0.7     Del if word ht lt xht x this
crunch_del_max_ht       3       Del if word ht gt xht x this
crunch_del_min_width    3       Del if word width lt xht x this
crunch_del_high_word    1.5     Del if word gt xht x this above bl
crunch_del_low_word     0.5     Del if word gt xht x this below bl
crunch_small_outlines_size      0.6     Small if lt xht x this
fixsp_small_outlines_size       0.28    Small if lt xht x this
superscript_worse_certainty     2        How many times worse certainty
does a superscript position glyph need to be for us to try classifying
it as a char with a different baseline?
superscript_bettered_certainty  0.97     What reduction in badness do
we think sufficient to choose a superscript over what we'd thought.
For example, a value of 0.6 means we want to reduce badness of
certainty by at least 40%
```

superscript_scaledown_ratio 0.4 A superscript scaled down more than this is unbelievably small. For example, 0.3 means we expect the font size to be no smaller than 30% of the text line font size.
subscript_max_y_top 0.5 Maximum top of a character measured as a multiple of x-height above the baseline for us to reconsider whether it's a subscript.
superscript_min_y_bottom 0.3 Minimum bottom of a character measured as a multiple of x-height above the baseline for us to reconsider whether it's a superscript.
suspect_rating_per_ch 999.9 Don't touch bad rating limit
suspect_accept_rating -999.9 Accept good rating limit
tessedit_lower_flip_hyphen 1.5 Aspect ratio dot/hyphen test
tessedit_upper_flip_hyphen 1.8 Aspect ratio dot/hyphen test
rej_whole_of_mostly_reject_word_fract 0.85 if >this fract
min_orientation_margin 7 Min acceptable orientation margin
textord_tabfind_vertical_text_ratio 0.5 Fraction of textlines deemed vertical to use vertical page mode
textord_tabfind_aligned_gap_fraction 0.75 Fraction of height used as a minimum gap for aligned blobs.
bestrate_pruning_factor 2 Multiplying factor of current best rate to prune other hypotheses
segment_reward_script 0.95 Score multipler for script consistency within a word. Being a 'reward' factor, it should be <= 1. Smaller value implies bigger reward.
segment_reward_chartype 0.97 Score multipler for char type consistency within a word.
segment_reward_ngram_best_choice 0.99 Score multipler for ngram permuter's best choice (only used in the Han script path).
heuristic_segcost_rating_base 1.25 base factor for adding segmentation cost into word rating.It's a multiplying factor, the larger the value above 1, the bigger the effect of segmentation cost.
heuristic_weight_rating 1 weight associated with char rating in combined cost ofstate
heuristic_weight_width 1000 weight associated with width evidence in combined cost of state
heuristic_weight_seamcut 0 weight associated with seam cut in combined cost of state
heuristic_max_char_wh_ratio 2 max char width-to-height ratio allowed in segmentation
segsearch_max_fixed_pitch_char_wh_ratio 2 Maximum character width-to-height ratio for fixed-pitch fonts
tosp_old_sp_kn_th_factor 2 Factor for defining space threshold in terms of space and kern sizes
tosp_threshold_bias1 0 how far between kern and space?
tosp_threshold_bias2 0 how far between kern and space?
tosp_narrow_fraction 0.3 Fract of xheight for narrow
tosp_narrow_aspect_ratio 0.48 narrow if w/h less than this
tosp_wide_fraction 0.52 Fract of xheight for wide
tosp_wide_aspect_ratio 0 wide if w/h less than this
tosp_fuzzy_space_factor 0.6 Fract of xheight for fuzz sp

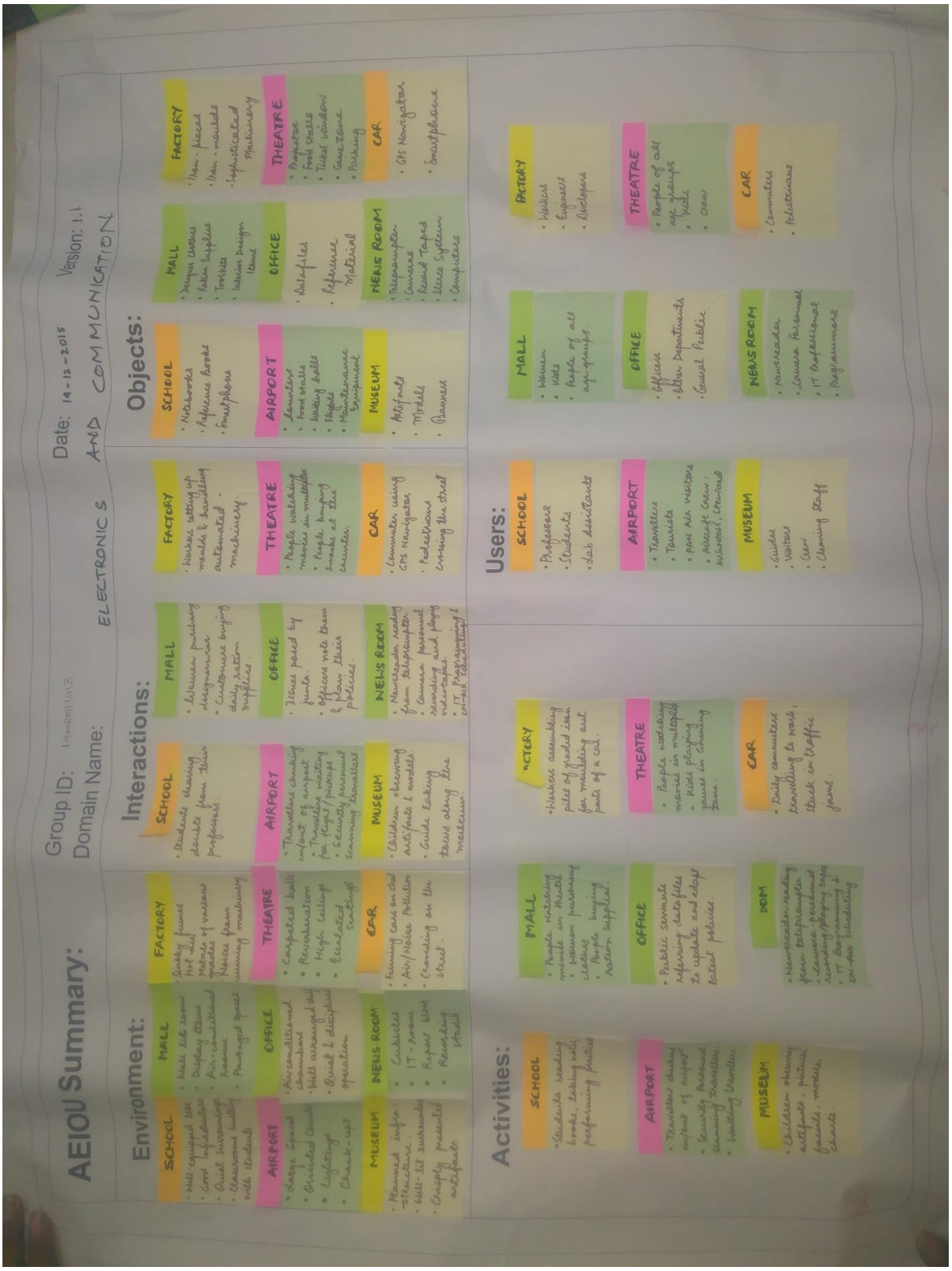
```
tosp_fuzzy_space_factor1        0.5     Fract of xheight for fuzz sp
tosp_fuzzy_space_factor2        0.72    Fract of xheight for fuzz sp
tosp_gap_factor 0.83    gap ratio to flip sp->kern
tosp_kern_gap_factor1   2       gap ratio to flip kern->sp
tosp_kern_gap_factor2   1.3     gap ratio to flip kern->sp
tosp_kern_gap_factor3   2.5     gap ratio to flip kern->sp
tosp_ignore_big_gaps    -1      xht multiplier
tosp_ignore_very_big_gaps       3.5     xht multiplier
tosp_rep_space  1.6     rep gap multiplier for space
tosp_enough_small_gaps   0.65     Fract of kerns reqd for isolated row
stats
tosp_table_kn_sp_ratio  2.25    Min difference of kn & sp in table
tosp_table_xht_sp_ratio 0.33    Expect spaces bigger than this
tosp_table_fuzzy_kn_sp_ratio    3       Fuzzy if less than this
tosp_fuzzy_kn_fraction  0.5     New fuzzy kn alg
tosp_fuzzy_sp_fraction  0.5     New fuzzy sp alg
tosp_min_sane_kn_sp     1.5     Don't trust spaces less than this time
kn
tosp_init_guess_kn_mult 2.2     Thresh guess - mult kn by this
tosp_init_guess_xht_mult          0.28     Thresh guess - mult xht by
this
tosp_max_sane_kn_thresh 5       Multiplier on kn to limit thresh
tosp_flip_caution         0        Don't autoflip kn to sp when large
separation
tosp_large_kerning      0.19    Limit use of xht gap with large kns
tosp_dont_fool_with_small_kerns -1       Limit use of xht gap with odd
small kns
tosp_near_lh_edge        0        Don't reduce box if the top left is
non blank
tosp_silly_kn_sp_gap    0.2     Don't let sp minus kn get too small
tosp_pass_wide_fuzz_sp_to_context        0.75    How wide fuzzies need
context
textord_blob_size_bigile        95      Percentile for large blobs
textord_noise_area_ratio         0.7      Fraction of bounding box for
noise
textord_blob_size_smallile      20      Percentile for small blobs
textord_initialx_ile    0.75    Ile of sizes for xheight guess
textord_initialasc_ile  0.9     Ile of sizes for xheight guess
textord_noise_sizelimit 0.5     Fraction of x for big t count
textord_noise_normratio 2       Dot to norm ratio for deletion
textord_noise_syfract   0.2     xh fract height error for norm blobs
textord_noise_sxfract   0.4     xh fract width error for norm blobs
textord_noise_hfract     0.015625          Height fraction to discard
outlines as speckle noise
textord_noise_rowratio  6       Dot to norm ratio for deletion
textord_blshift_maxshift        0       Max baseline shift
textord_blshift_xfraction       9.99    Min size of baseline shift
```

Design: Analysis, Design Methodology and Implementation Strategy

1. **Observation Record Sheets:**
    - **AEIOU Summary**
    - **Activities**
    - **Environment**
    - **Interactions**
    - **Objects**
    - **Users**
    - **Flow Diagram**

2. **Canvases:**
    - **Empathy**
    - **Ideation**
    - **Product Development**

AEIOU Summary



AEIOU framework:

Group id: 1400801110 13 Date: Sheet No: 1

Activities

Project Name : PERSONAL INFO ASSISTANT

General impressions / Observations

School : Students are reading books, taking notes, performing practicals, extra time spent in canteen.

Factory : Workers are assembling piles of iron of different grades for moulding out parts of a car.

Office : Public servants are referring piles of data files to update and apply latest government policies.

Museum : Children are observing artifacts, pictures, fossils- proofs of life from the Past.

Car : A daily commuter, driving to work, is stuck in a traffic jam.

Elements, features and special notes

ooks, notebooks, trainer kits, lab equipments

Moulds, reference manuals, grade samples, robotic arms

les, folders, catalogues, reference manuals, policy guides

PS Navigators, Stereo system, mobile phone, office baggage

ssils, Showcases, charts, models, busts (statues),

nreferenced display units.

Sketch/photo- Summary of activities

EIOU framework:

Environment

Group id: 14008011013 Date: Sheet No: 2

Project Name : PERSONAL INFO ASSISTANT

General impressions / Observations

(style, materials & atmosphere)

ool : Well equipped labs, good infrastructure, classrooms bustling with students, quiet surroundings

tory : Smoky fumes, hot air, metals of different grades, noises from running machines

ce : Air conditioned chambers, chairs & desks, quiet and disciplined operation.

eum: Planned infrastructure, well lit surroundings, well presented artifacts, neat and clean dorms.

: Fuming cars, air/noise pollution, crowding on the street.

Floor plan

Optimized space usage, required stationery and books.

Rooms at maintained temperatures, fixed moulds, conveyor, robotic arms for mechanization.

Optimized data storage, space for one to one interaction, cool temperatures.

Infrastructure which suits/supports longevity of artifacts, proper amenities for visitors.

Maintained cars, clear roads, streetlights, horns.

ements, features and special notes

rastructure, lab equipements, display boards

lated rooms/moulds, heavy machinery, hot and uted air, burning fuels.

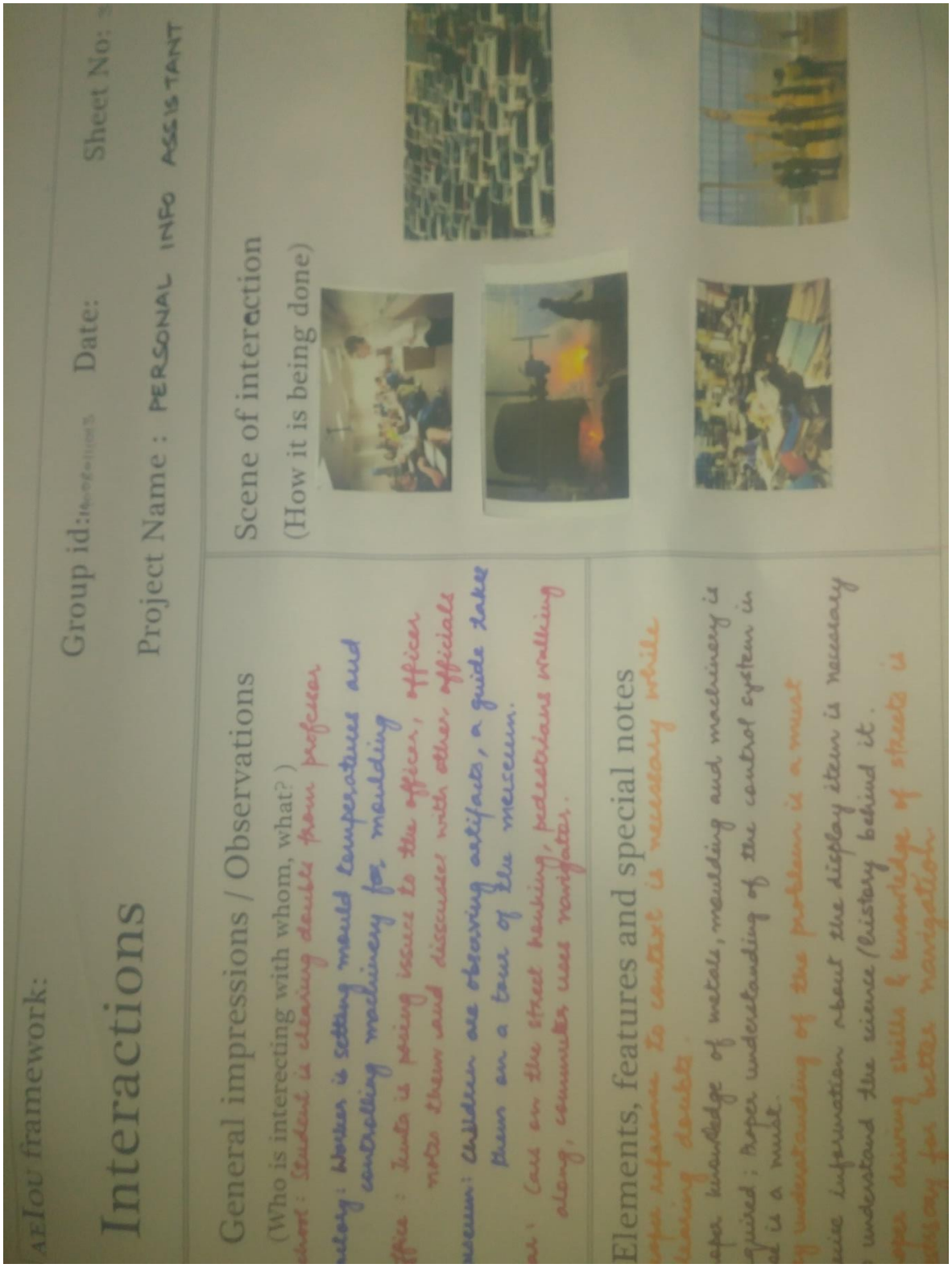
matic arrangement of data files, cupboards collection of previous sheets.

maintained, use of variety elements to create desired ience, lighting creates aesthetic display.

tic movement of cars, difficult navigation, sions due to unplanned travelling routes.

Scene

Interactions



AEIOU framework:

Group id: Date: Sheet No:

Interactions

Project Name : PERSONAL INFO ASSISTANT

General impressions / Observations

(Who is interecting with whom, what?)

Student : Student is clearing doubts from professor

Lecture : Lecture is setting mould temperatures and controlling machinery for moulding

Office : Lecture is posing issues to the officer, officer notes them and discusses with other officials

Museum : Children are observing artifacts, a guide takes them on a tour of the museum.

Bus : Cars on the street honking, pedestrians walking along, commuters use navigation.

Scene of interaction

(How it is being done)

Elements, features and special notes

Lecture information about the display item is necessary

understood the science/history behind it.

Lecture during skills & knowledge of students is necessary for better navigation.

Lecture information is essential to control is necessary while clearing doubts.

Lecture knowledge of metals, moulding and machinery is required : Proper understanding of the control system in lecture is a must.

Lecture understanding of the problem is a must

AEIOU framework:

Group id: 140080113013 Date: Sheet No: 4

Objects

Project Name : PERSONAL INFO ASSISTANT

General impressions / Observations

(What components are involved?)

School : Notebook, reference books, dictionary, smartphone

Factory : Moulds, sophisticated machinery, iron pieces, robotic arms

Office : Datafiles, Policy manuals, reference material

Museum : Artifacts, models, banners

Car : GPS Navigator, smartphone

Elements, features and special notes

(How objects are relating to the activities?)

Notes are taken in classes & from reference books

Sophisticated machinery is used to create iron moulds from desired grade.

Datafiles & policy manuals are studied before policy making.

Banners give an idea about the item on display.

GPS enables choice of better route in case of a traffic jam.

Inventory of key objects

SCHOOL :
- NOTEBOOKS
- REFERENCE BOOKS
- SMARTPHONE

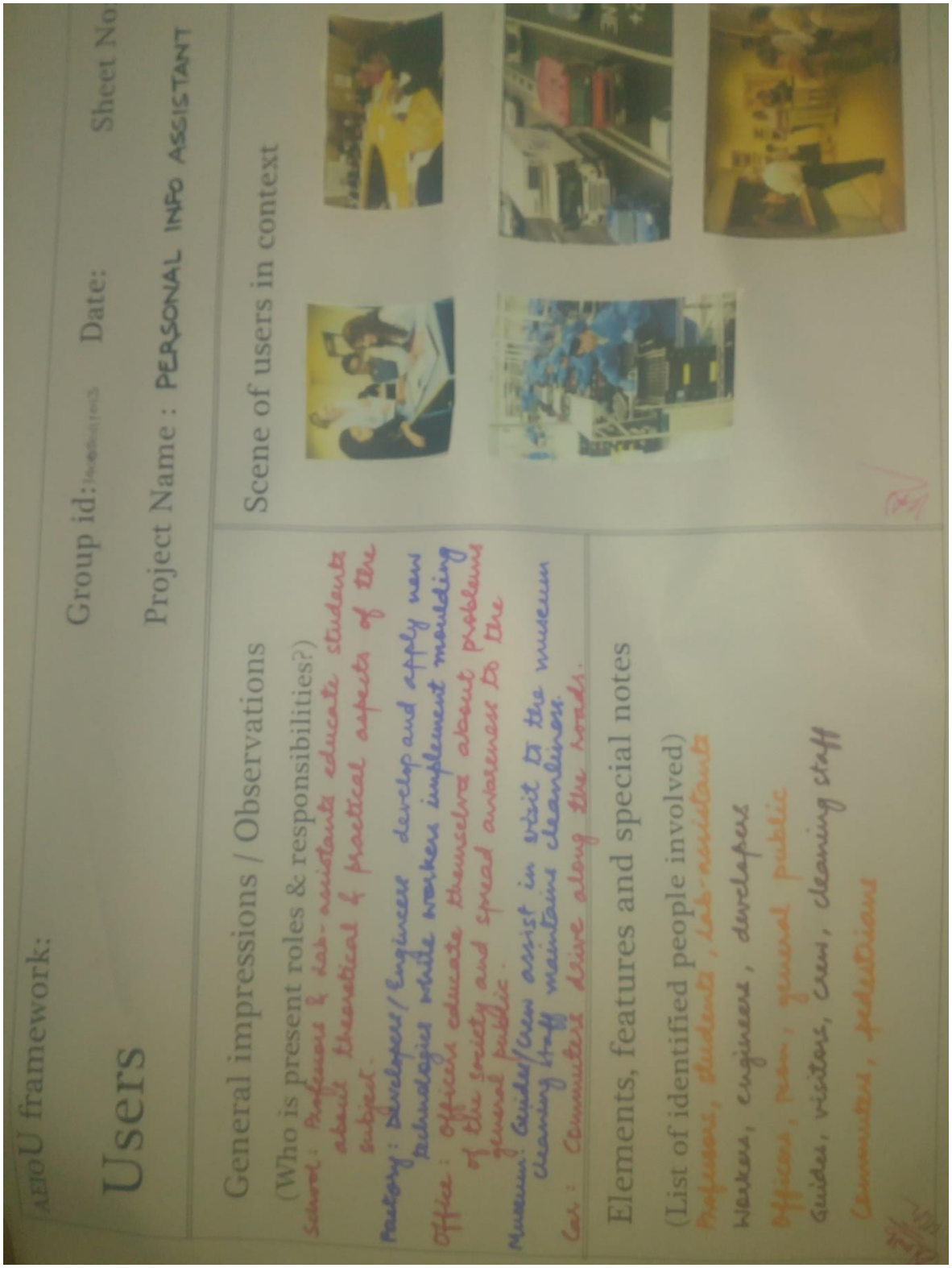
FACTORY :
- IRON PIECES
- IRON MOULDS
- SOPHISTICATED MACHINERY

OFFICE :
- DATA FILES
- REFERENCE MATERIAL

MUSEUM :
- ARTIFACTS
- MODELS
- BANNERS

CAR :
- GPS NAVIGATOR
- SMARTPHONE

Users



AEIOU framework:

Users

Group id: Date: Sheet No

Project Name : PERSONAL INFO ASSISTANT

General impressions / Observations

(Who is present roles & responsibilities?)

School: Professors & lab-assistants educate students about theoretical & practical aspects of the subject.

Pottery: Developers/Engineers develop and apply new technologies while workers implement moulding

Office: Officers educate themselves about problems of the society and spread awareness to the general public.

Museum: Guides/crew assist in visit to the museum cleaning staff maintains cleanliness.

Lab: Computers have along the roads.

Scene of users in context

Elements, features and special notes

(List of identified people involved)

Professors, students, lab-assistants

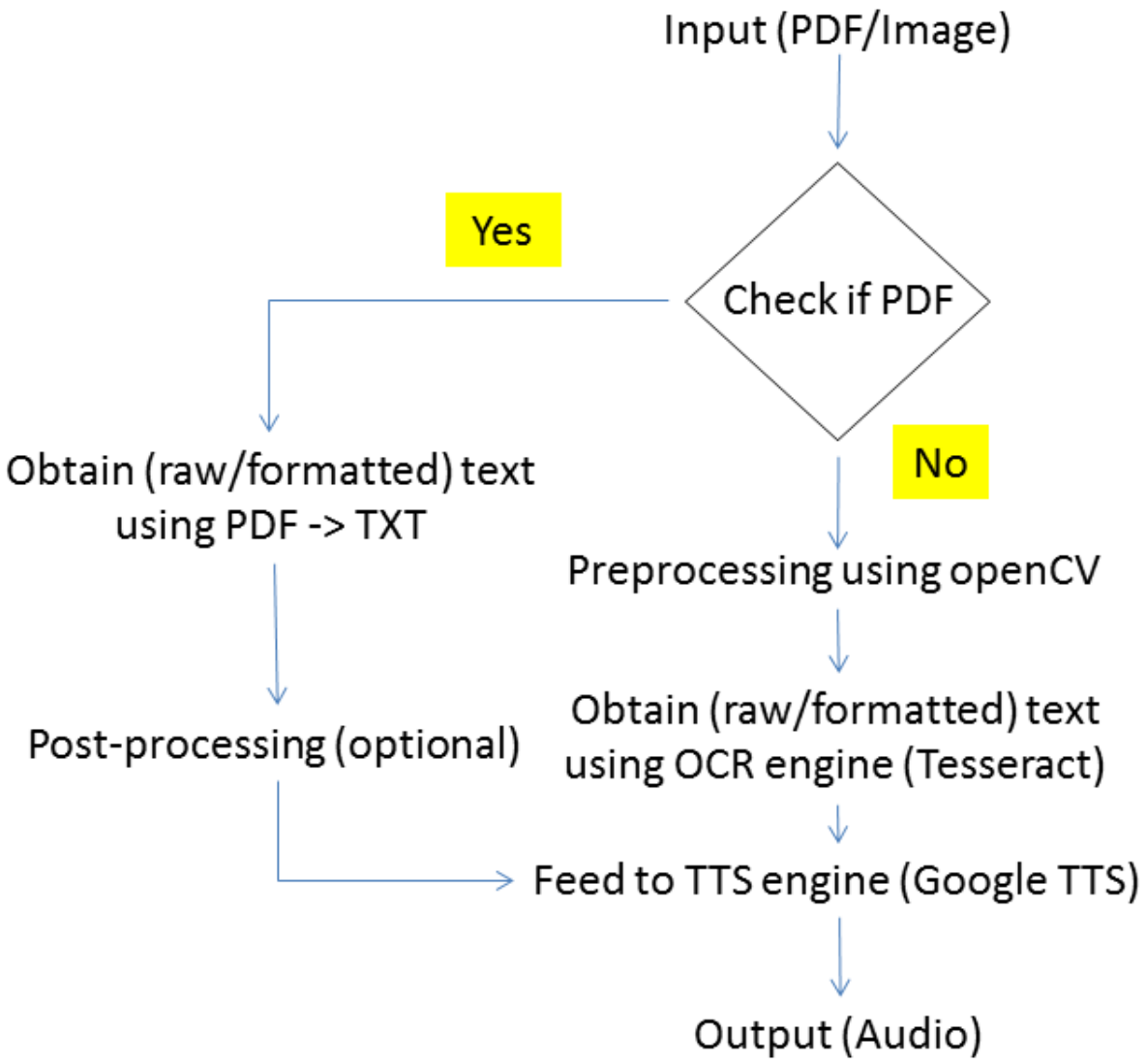
Workers, engineers, developers

Officers, men, general public

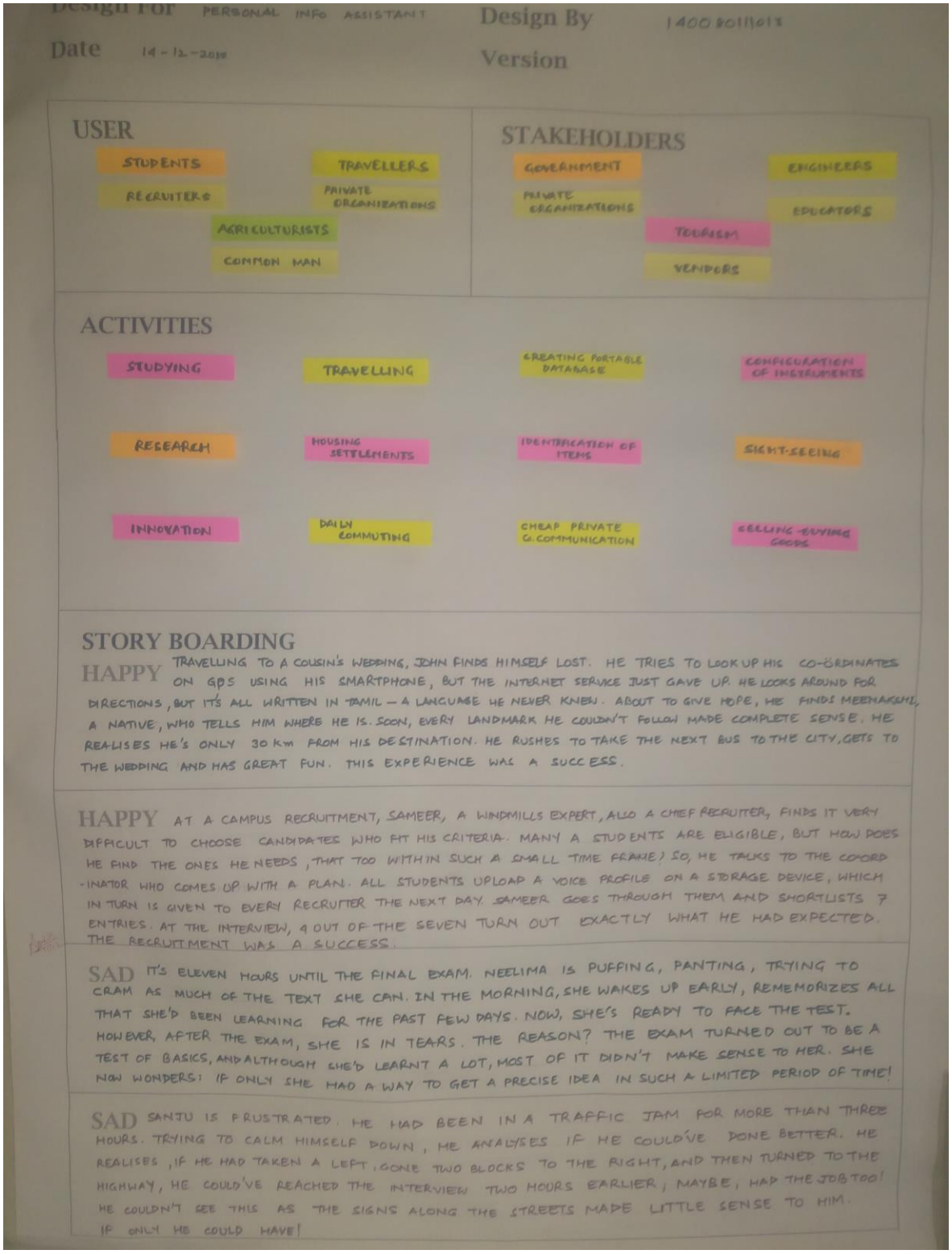
Guides, visitors, crew, cleaning staff

Computers, pedestrians

Flow Diagram



Canvas: Empathy



Design For PERSONAL INFO ASSISTANT

Design By 1400801101013

Date 14-12-2018

Version

USER

STUDENTS

TRAVELLERS

RECRUITERS

PRIVATE ORGANIZATIONS

AGRICULTURISTS

COMMON MAN

STAKEHOLDERS

GOVERNMENT

ENGINEERS

PRIVATE ORGANIZATIONS

EDUCATORS

TOURISM

VENDORS

ACTIVITIES

STUDYING

TRAVELLING

CREATING PORTABLE DATABASE

CONFIGURATION OF INSTRUMENTS

RESEARCH

HOUSING SETTLEMENTS

IDENTIFICATION OF ITEMS

SIGHT-SEEING

INNOVATION

DAILY COMMUTING

CHEAP PRIVATE C. COMMUNICATION

SELLING-BUYING GOODS

STORY BOARDING

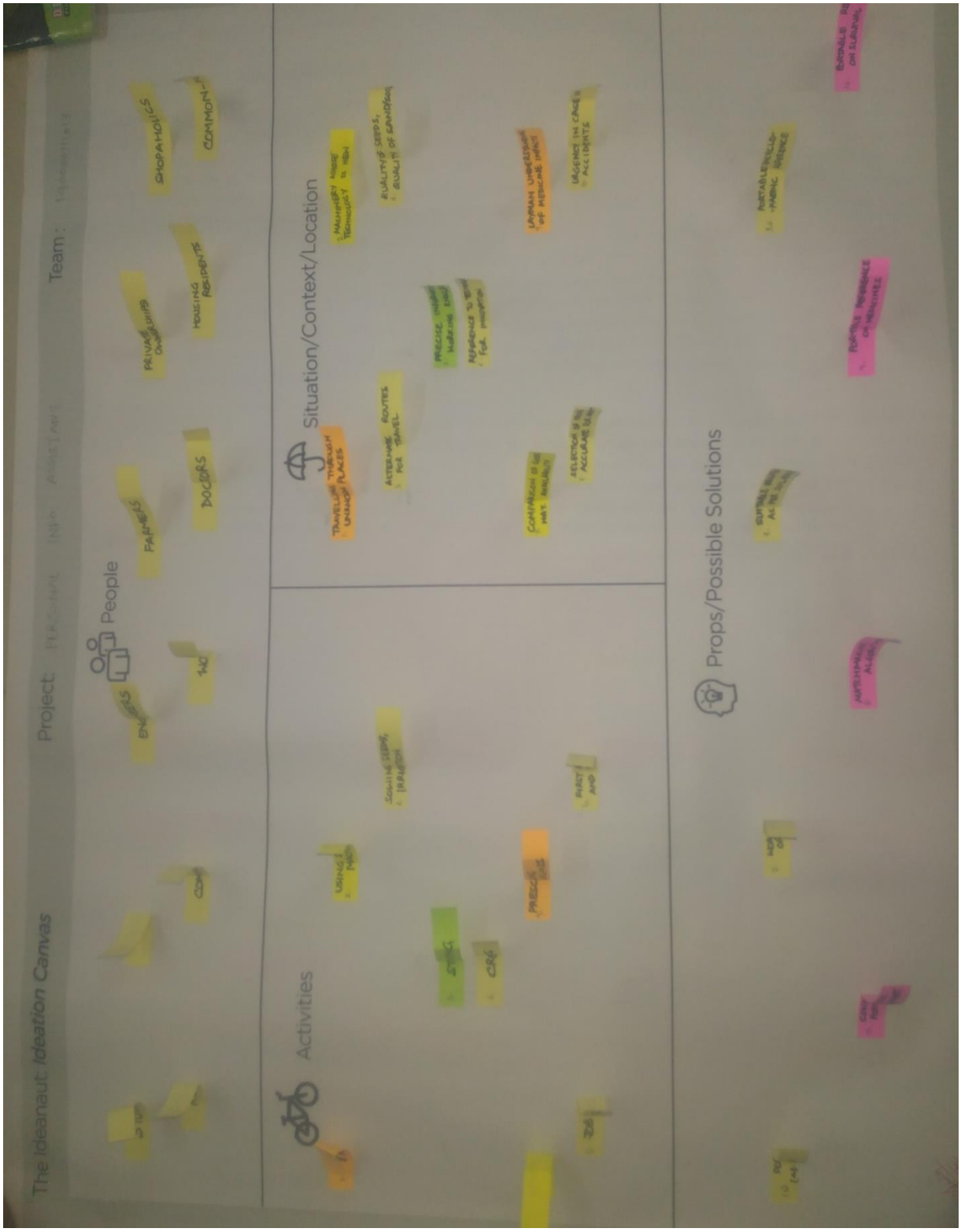
HAPPY TRAVELLING TO A COUSIN'S WEDDING, JOHN FINDS HIMSELF LOST. HE TRIES TO LOOK UP HIS CO-ÖRDINATES ON GPS USING HIS SMARTPHONE, BUT THE INTERNET SERVICE JUST GAVE UP. HE LOOKS AROUND FOR DIRECTIONS, BUT IT'S ALL WRITTEN IN TAMIL — A LANGUAGE HE NEVER KNEW. ABOUT TO GIVE HOPE, HE FINDS MEENAKSHI, A NATIVE, WHO TELLS HIM WHERE HE IS. SOON, EVERY LANDMARK HE COULDN'T FOLLOW MADE COMPLETE SENSE. HE REALISES HE'S ONLY 30 Km FROM HIS DESTINATION. HE RUSHES TO TAKE THE NEXT BUS TO THE CITY, GETS TO THE WEDDING AND HAS GREAT FUN. THIS EXPERIENCE WAS A SUCCESS.

HAPPY AT A CAMPUS RECRUITMENT, SAMEER, A WINDMILLS EXPERT, ALSO A CHIEF RECRUITER, FINDS IT VERY DIFFICULT TO CHOOSE CANDIDATES WHO FIT HIS CRITERIA. MANY A STUDENTS ARE ELIGIBLE, BUT HOW DOES HE FIND THE ONES HE NEEDS, THAT TOO WITHIN SUCH A SMALL TIME FRAME? SO, HE TALKS TO THE CO-ORD-INATOR WHO COMES UP WITH A PLAN. ALL STUDENTS UPLOAD A VOICE PROFILE ON A STORAGE DEVICE, WHICH IN TURN IS GIVEN TO EVERY RECRUITER THE NEXT DAY. SAMEER GOES THROUGH THEM AND SHORTLISTS 7 ENTRIES. AT THE INTERVIEW, 4 OUT OF THE SEVEN TURN OUT EXACTLY WHAT HE HAD EXPECTED. THE RECRUITMENT WAS A SUCCESS.

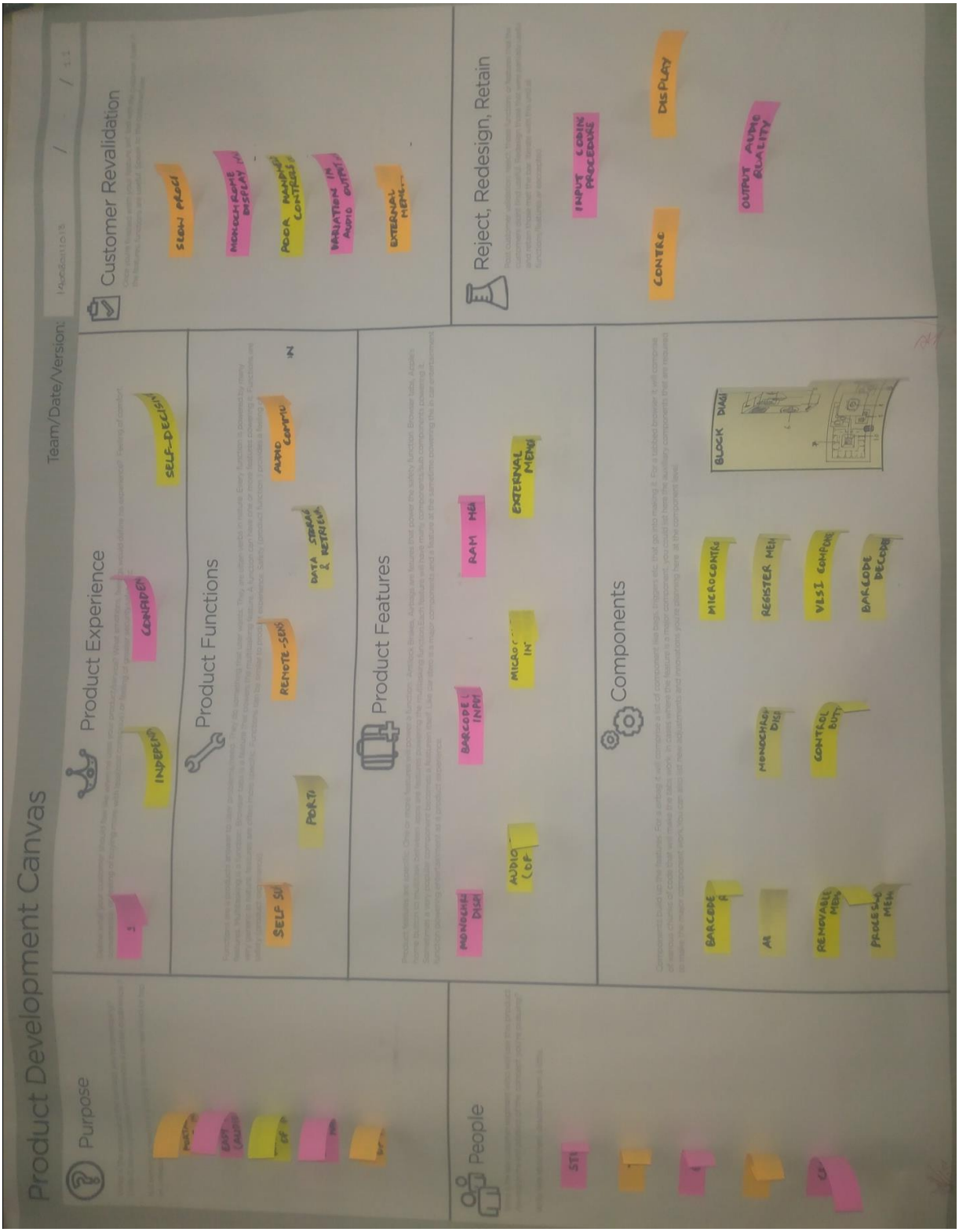
SAD IT'S ELEVEN HOURS UNTIL THE FINAL EXAM. NEELIMA IS PUFFING, PANTING, TRYING TO CRAM AS MUCH OF THE TEXT SHE CAN. IN THE MORNING, SHE WAKES UP EARLY, REMEMORIZES ALL THAT SHE'D BEEN LEARNING FOR THE PAST FEW DAYS. NOW, SHE'S READY TO FACE THE TEST. HOWEVER, AFTER THE EXAM, SHE IS IN TEARS. THE REASON? THE EXAM TURNED OUT TO BE A TEST OF BASICS, AND ALTHOUGH SHE'D LEARNT A LOT, MOST OF IT DIDN'T MAKE SENSE TO HER. SHE NOW WONDERS: IF ONLY SHE HAD A WAY TO GET A PRECISE IDEA IN SUCH A LIMITED PERIOD OF TIME!

SAD SANJU IS FRUSTRATED. HE HAD BEEN IN A TRAFFIC JAM FOR MORE THAN THREE HOURS. TRYING TO CALM HIMSELF DOWN, HE ANALYSES IF HE COULD'VE DONE BETTER. HE REALISES, IF HE HAD TAKEN A LEFT, GONE TWO BLOCKS TO THE RIGHT, AND THEN TURNED TO THE HIGHWAY, HE COULD'VE REACHED THE INTERVIEW TWO HOURS EARLIER; MAYBE, HAD THE JOB TOO! HE COULDN'T SEE THIS AS THE SIGNS ALONG THE STREETS MADE LITTLE SENSE TO HIM. IF ONLY HE COULD HAVE!

Canvas: Ideation



Canvas: Product Development



Implementation

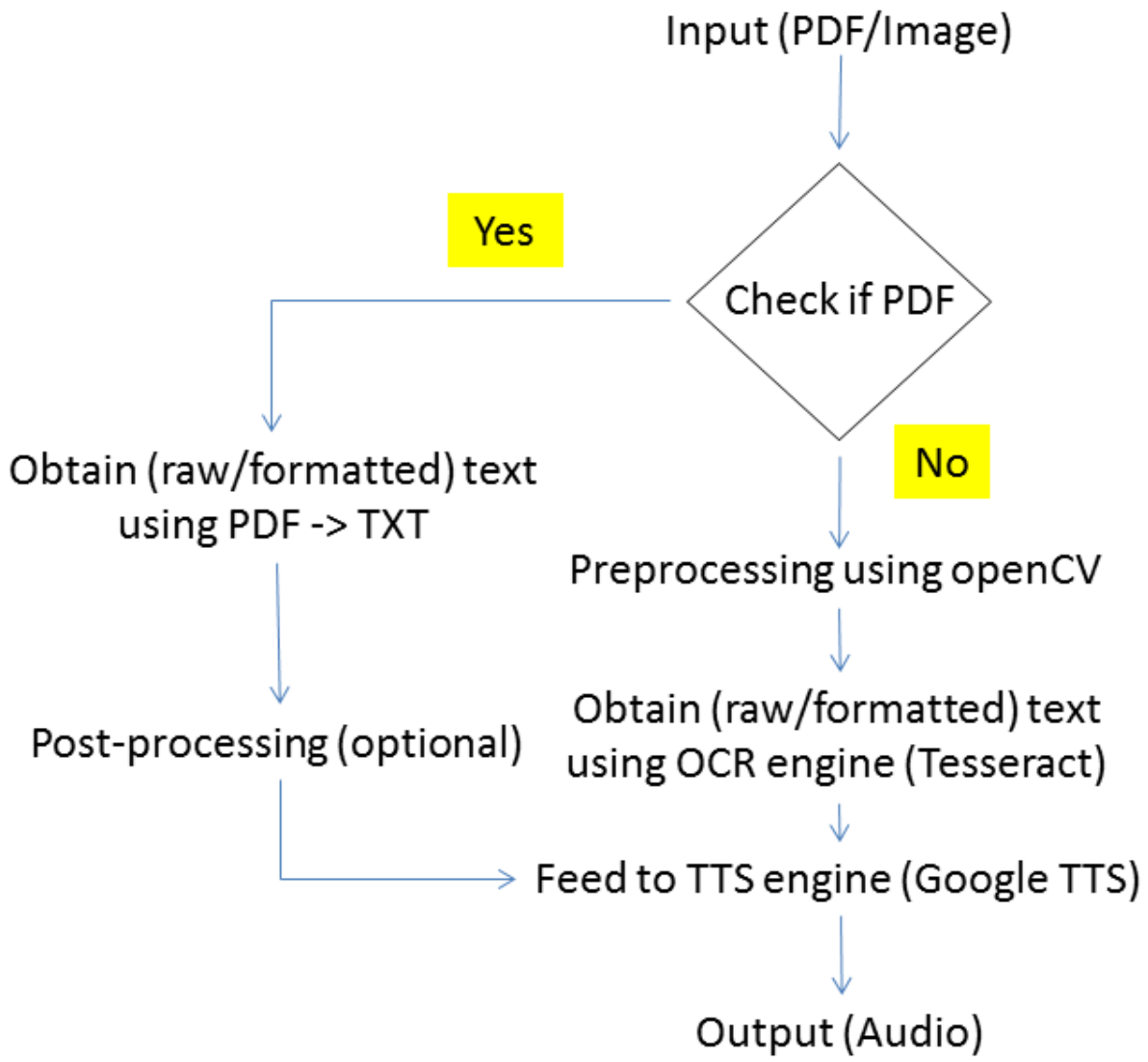
1. **Block Diagram**

2. **Progress Remarks**

3. **Softwares**

4. **Source Code Listing**
    - *pytesseract.py*
    - *ocr_pia_1.py*
    - *ocr_pia_2.py*
    - *thresholding_adaptive_naive.py*
    - *thresholding_global_naive.py*
    - *thresholding_global_naive2.py*
    - *thresholding_global_otsu.py*
    - *thresholding_global_otsu2.py*

Block Diagram



Progress Remarks

- **OCR Pre-processing**
  The first step towards a mobile-platform implementation for our project was to identify the key areas to be resolved to improve text recognition.
  We identified these areas to be:

1. Choice of lossless compression image format (eg. PNG)
2. Adaptive Binarization of image
3. Fragmentation
4. Filter Application

- **Source Code**
  The source-code will be made available (alongwith necessary datasets) on my Github page:
  http://www.github.com/CRT13/Projects/

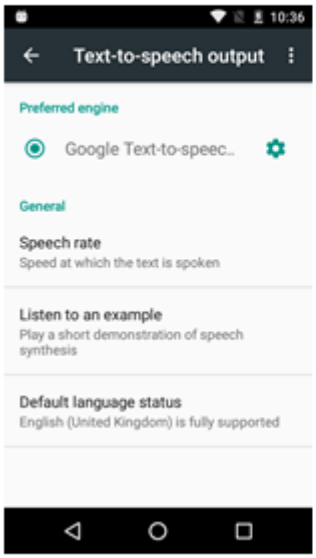
Softwares

QPython3 - Python3 for Android

Tesseract OCR

SL4A – Scripting Layer for Android

Google Text-to-speech



Source Code Listing

1. **pytesseract.py**
   Python3 wrapper for interfacing Tesseract-OCR with OpenCV3 library.
2. **ocr_pia_1.py**
   Online OCR (uses WeOCR server)
3. **ocr_pia_2.py**
   Offline OCR (uses "pytesseract" module)
4. **thresholding_adaptive_naive.py**
   Adaptive thresholding using OpenCV
5. **thresholding_global_naive.py**
   Global thresholding using OpenCV
6. **thresholding_global_naive2.py**
   Global thresholding using OpenCV
7. **thresholding_global_otsu.py**
   Global thresholding using Otsu's Binarization in OpenCV
8. **thresholding_global_otsu2.py**
   Global thresholding using Otsu's Binarization in OpenCV

pytesseract.py

```
#!/usr/bin/env python
##############################################################################
try:
    import Image
except ImportError:
    from PIL import Image

import os
import sys
import subprocess
import tempfile
import shlex

# CHANGE THIS IF TESSERACT IS NOT IN YOUR PATH, OR IS NAMED DIFFERENTLY
tesseract_cmd = 'tesseract'

__all__ = ['image_to_string']

def run_tesseract(input_filename, output_filename_base, lang=None,
boxes=False,config=None):
    '''
    runs the command:
        `tesseract_cmd` `input_filename` `output_filename_base`
    returns the exit status of tesseract, as well as tesseract's stderr
output
    '''
    command = [tesseract_cmd, input_filename, output_filename_base]
    if lang is not None:
        command += ['-l', lang]
    if boxes:
        command += ['batch.nochop', 'makebox']
    if config:
        command += shlex.split(config)

    proc = subprocess.Popen(command, stderr=subprocess.PIPE)
    status = proc.wait()
    error_string = proc.stderr.read()
    proc.stderr.close()
    return status, error_string

def cleanup(filename):
    ''' tries to remove the given filename. Ignores non-existent files '''
    try:
        os.remove(filename)
    except OSError:
        pass

def get_errors(error_string):
    '''
    returns all lines in the error_string that start with the string "error"
    '''
    error_string = error_string.decode('utf-8')
    lines = error_string.splitlines()
    error_lines = tuple(line for line in lines if line.find(u'Error') >= 0)
```

```
    if len(error_lines) > 0:
        return u'\n'.join(error_lines)
    else:
        return error_string.strip()


def tempnam():
    ''' returns a temporary file-name '''
    tmpfile = tempfile.NamedTemporaryFile(prefix="tess_")
    return tmpfile.name

class TesseractError(Exception):
    def __init__(self, status, message):
        self.status = status
        self.message = message
        self.args = (status, message)

    def image_to_string(image, lang=None, boxes=False, config=None):
    '''
    Runs tesseract on the specified image. First, the image is written to
disk,
    and then the tesseract command is run on the image. Tesseract's result is
    read, and the temporary files are erased.
    Also supports boxes and config:

    if boxes=True
        "batch.nochop makebox" gets added to the tesseract call

    if config is set, the config gets appended to the command.
        ex: config="-psm 6"
    '''
        if len(image.split()) == 4:
        # In case we have 4 channels, lets discard the Alpha.
        # Kind of a hack, should fix in the future some time.
            r, g, b, a = image.split()
            image = Image.merge("RGB", (r, g, b))

        input_file_name = '%s.bmp' % tempnam()
        output_file_name_base = tempnam()
        if not boxes:
            output_file_name = '%s.txt' % output_file_name_base
        else:
            output_file_name = '%s.box' % output_file_name_base
        try:
            image.save(input_file_name)
            status, error_string = run_tesseract(input_file_name,
                                                 output_file_name_base,
                                                 lang=lang,
                                                 boxes=boxes,
                                                 config=config)
            if status:
                errors = get_errors(error_string)
                raise TesseractError(status, errors)
            f = open(output_file_name, 'rb')
            try:
```

```
                return f.read().decode('utf-8').strip()
            finally:
                f.close()

    finally:
        cleanup(input_file_name)
        cleanup(output_file_name)


def main():
    if len(sys.argv) == 2:
        filename = sys.argv[1]
        try:
            image = Image.open(filename)
            if len(image.split()) == 4:
                # In case we have 4 channels, lets discard the Alpha.
                # Kind of a hack, should fix in the future some time.
                r, g, b, a = image.split()
                image = Image.merge("RGB", (r, g, b))
        except IOError:
            sys.stderr.write('ERROR: Could not open file "%s"\n' % filename)
            exit(1)
        print(image_to_string(image))
    elif len(sys.argv) == 4 and sys.argv[1] == '-l':
        lang = sys.argv[2]
        filename = sys.argv[3]
        try:
            image = Image.open(filename)
        except IOError:
            sys.stderr.write('ERROR: Could not open file "%s"\n' % filename)
            exit(1)
        print(image_to_string(image, lang=lang))
    else:
        sys.stderr.write('Usage: python pytesseract.py [-l lang]
input_file\n')
        exit(2)
##############################################################################

if __name__ == '__main__':
    main()
```

ocr_pia_1.py

```
import http.client, mimetypes, android, sys, sl4a

droid = android.Android()

def post_multipart(host, selector, fields, files):
    content_type, body = encode_multipart_formdata(fields, files)
    h = httplib.HTTP(host)
    h.putrequest('POST', selector)
    h.putheader('content-type', content_type)
    h.putheader('content-length', str(len(body)))
    h.endheaders()
    h.send(body)
    errcode, errmsg, headers = h.getreply()
    return h.file.read()

def encode_multipart_formdata(fields, files):
    BOUNDARY = '----------ThIs_Is_tHe_bouNdaRY_$'
    CRLF = '\r\n'
    L = []
    for (key, value) in fields:
        L.append('--' + BOUNDARY)
        L.append('Content-Disposition: form-data; name="%s"' % key)
        L.append('')
        L.append(value)
    for (key, filename, value) in files:
        L.append('--' + BOUNDARY)
        L.append('Content-Disposition: form-data; name="%s"; filename="%s"' %
(key, filename))
        L.append('Content-Type: %s' % get_content_type(filename))
        L.append('')
        L.append(value)
    L.append('--' + BOUNDARY + '--')
    L.append('')
    body = CRLF.join(L)
    content_type = 'multipart/form-data; boundary=%s' % BOUNDARY
    return content_type, body

def get_content_type(filename):
    return mimetypes.guess_type(filename)[0] or 'application/octet-stream'

##########################################################################
###
""" Open html file after writing it, my info for the main app screen """
im2cote = open('/sdcard/im2cote.html','w')

im2cote.write('\n'.join(['<html><head></head><div
style="width:250px;height:370px;border:5px solid grey;"><body
bgcolor="#000000"><h><p>...</p>.<font color="#FFFFFF"><font
size="4">   SeeSay - by Dave
Cote</h1><p>         <font
color="#FFFFFF"><font size="2">e: info@davcote.com</font></p><p><font
color="#000000">.....</font></p><p><font color="#FFFFFF"> <font
size="3">Please zoom in to take picture</font></p><font
```

```
color="#FFFFFF"> <font size="3">otherwise the picture file
will</font></p><p><font colour="#FFFFFF"> <font size="3">be too large to
decode . . . . . .</font></p></body></html></div>']))

im2cote.close()
droid.webViewShow('file:///sdcard/im2cote.html')




""" Capture image using camera """
droid.cameraInteractiveCapturePicture('/sdcard/jpeg.jpg')

                                    """ Optional """

""" Resize new jpeg by extracting it's thumbnail with EXIF.py and resaving
"""
#fileTHUMB = open("jpeg.jpg",'rb')
#tags = EXIF.process_file(fileTHUMB)
""" Save the thumbnail into the jpg format file"""
#fileTHUMB = open("jpeg.jpg",'wb')
#fileTHUMB.write(tags["JPEGThumbnail"])
#fileTHUMB.close()

""" Send image to cloud """
host = 'appsv.ocrgrid.org'
selector = '/cgi-bin/weocr/submit_tesseract.cgi'
fields = [('outputencoding', 'utf-8'), ('outputformat', 'txt')]
with open('/sdcard/jpeg.jpg', 'rb') as jpeg:
    files = [('userfile', 'jpeg.jpg', jpeg.read())]
response = post_multipart(host, selector, fields, files)

""" Cleanup garbage:  WORKS!!! """
def ExtractAlphaNumeric(InputString):
   from string import ascii_letters,digits
   return "".join([ch for ch in InputString if ch in (ascii_letters+digits+'
'+'.'+','+'\n'+'!'+'?'+'@'+'$'+'%'+'&')])

response = ExtractAlphaNumeric(response)
""" Speak response """
droid.ttsSpeak(response)
```

ocr_pia_2.py

```
from PIL import Image
import pytesseract
import os,sys
""" Cleanup Garbage Values """
def ocr_cleanup(s):
    s = ''.join(filter(lambda x: ord(x)<128,s))
    garbage = '[]{}<>\\\n*#*&^@:'
    for x in range(0,len(garbage)):
        s = s.replace(garbage[x],' ')
    return s
if __name__ == '__main__':
    #in_img = sys.argv[1]
    in_img = 'C:\\Users\\CRT13\\Desktop\\ocrPIA\\images\\' + input('Image to
be processed: ')
    try:
        out_ocr =
pytesseract.image_to_string(Image.open(in_img),lang='eng')#,boxes=False,confi
g='CT13-Test1')
        out_ocr = ocr_cleanup(out_ocr)
        print(out_ocr)
    except IOError:
        print('Unable to find',in_img)
```

thresholding_adaptive_naive.py

```
""" Python3: Adaptive Thresholding using OpenCV """
import numpy as np
from matplotlib import pyplot as plt
import cv2
# Load I/P image
in_img = 'C:\\Users\\CRT13\\Desktop\\Image
Processing\\OpenCV\\images\\'+input('Choose image to process: ')
img = cv2.imread(in_img,0)    # Load image in grayscale

# Thresholding Algorithms
r1,t1 = cv2.threshold(img,127,255,cv2.THRESH_BINARY)
t2 = cv2.adaptiveThreshold(img,255,cv2.ADAPTIVE_THRESH_MEAN_C,
            cv2.THRESH_BINARY,11,2)
t3 = cv2.adaptiveThreshold(img,255,cv2.ADAPTIVE_THRESH_GAUSSIAN_C,
            cv2.THRESH_BINARY,11,2)
# Image-display Routines
images = [img,t1,t2,t3]
titles = ['Original Image','Global Thresholding (v=127)','Adaptive-Mean
Thresholding','Adaptive-Gaussian Thresholding']
for i in range(4):
    plt.subplot(2,2,i+1),plt.imshow(images[i],'gray')
    plt.title(titles[i])
    plt.xticks([]),plt.yticks([])
plt.show()
```

thresholding_global_naive.py

```
""" Python3: Global Thresholding using OpenCV """
import numpy as np
from matplotlib import pyplot as plt
import cv2
# Load I/P image
in_img = 'C:\\Users\\CRT13\\Desktop\\Image
Processing\\OpenCV\\images\\'+input('Choose image to process: ')
img = cv2.imread(in_img,0)  # Load image in grayscale
# Simple Thresholding Algorithms
r1,t1 = cv2.threshold(img,127,255,cv2.THRESH_BINARY)
r2,t2 = cv2.threshold(img,127,255,cv2.THRESH_BINARY_INV)
r3,t3 = cv2.threshold(img,127,255,cv2.THRESH_TRUNC)
r4,t4 = cv2.threshold(img,127,255,cv2.THRESH_TOZERO)
r5,t5 = cv2.threshold(img,127,255,cv2.THRESH_TOZERO_INV)
# Image-display Routines
images = [img,t1,t2,t3,t4,t5]
titles = ['Original
Image','BINARY','BINARY_INV','TRUNC','TOZERO','TOZERO_INV']
for i in range(6):
    plt.subplot(2,3,i+1),plt.imshow(images[i],'gray')
    plt.title(titles[i])
    plt.xticks([]),plt.yticks([])
plt.show()
```

thresholding_global_naive2.py

```
""" Python3: Global Thresholding using OpenCV """
import numpy as np
from matplotlib import pyplot as plt
import cv2
# Load I/P image
in_img = 'C:\\Users\\CRT13\\Desktop\\Image
Processing\\OpenCV\\images\\'+input('Choose image to process: ')
img_org = cv2.imread(in_img,0)   # Load image in grayscale
img = cv2.GaussianBlur(img_org,(5,5),0)     # Gaussian Filtering
# Simple Thresholding Algorithms
r1,t1 = cv2.threshold(img,127,255,cv2.THRESH_BINARY)
r2,t2 = cv2.threshold(img,127,255,cv2.THRESH_BINARY_INV)
r3,t3 = cv2.threshold(img,127,255,cv2.THRESH_TRUNC)
r4,t4 = cv2.threshold(img,127,255,cv2.THRESH_TOZERO)
r5,t5 = cv2.threshold(img,127,255,cv2.THRESH_TOZERO_INV)
# Image-display Routines
images = [img,t1,t2,t3,t4,t5]
titles = ['Original
Image','BINARY','BINARY_INV','TRUNC','TOZERO','TOZERO_INV']
for i in range(6):
    plt.subplot(2,3,i+1),plt.imshow(images[i],'gray')
    plt.title(titles[i])
    plt.xticks([]),plt.yticks([])
plt.show()
```

thresholding_global_otsu.py

```
""" Python3: Global Thresholding (Otsu's Method) using OpenCV """
import numpy as np
from matplotlib import pyplot as plt
import cv2

in_img = 'C:\\Users\\CRT13\\Desktop\\Image
Processing\\OpenCV\\images\\'+input('Choose image to process: ')
img = cv2.imread(in_img,0)
img_blur = cv2.GaussianBlur(img,(5,5),0)
r1,t1 = cv2.threshold(img,127,255,cv2.THRESH_BINARY)
r2,t2 = cv2.threshold(img,0,255,cv2.THRESH_BINARY+cv2.THRESH_OTSU)
r3,t3 =
cv2.threshold(img_blur,0,255,cv2.THRESH_BINARY+cv2.THRESH_OTSU)
images = [img,0,t1,
          img,0,t2,
          img_blur,0,t3]
titles = ['Original Image','Histogram','Global Thresholding',
          'Original Image','Histogram','Otsu\'s Thresholding',
          'Gaussian-filtered Image','Histogram','Otsu\'s
Thresholding']
for i in range(3):
    plt.subplot(3,3,i*3+1),plt.imshow(images[i*3],'gray')
    plt.title(titles[i*3]),plt.xticks([]),plt.yticks([])
    plt.subplot(3,3,i*3+2),plt.hist(images[i*3].ravel(),256)
    plt.title(titles[i*3+1]),plt.xticks([]),plt.yticks([])
    plt.subplot(3,3,i*3+3),plt.imshow(images[i*3+2],'gray')
    plt.title(titles[i*3+2]),plt.xticks([]),plt.yticks([])
plt.show()
```

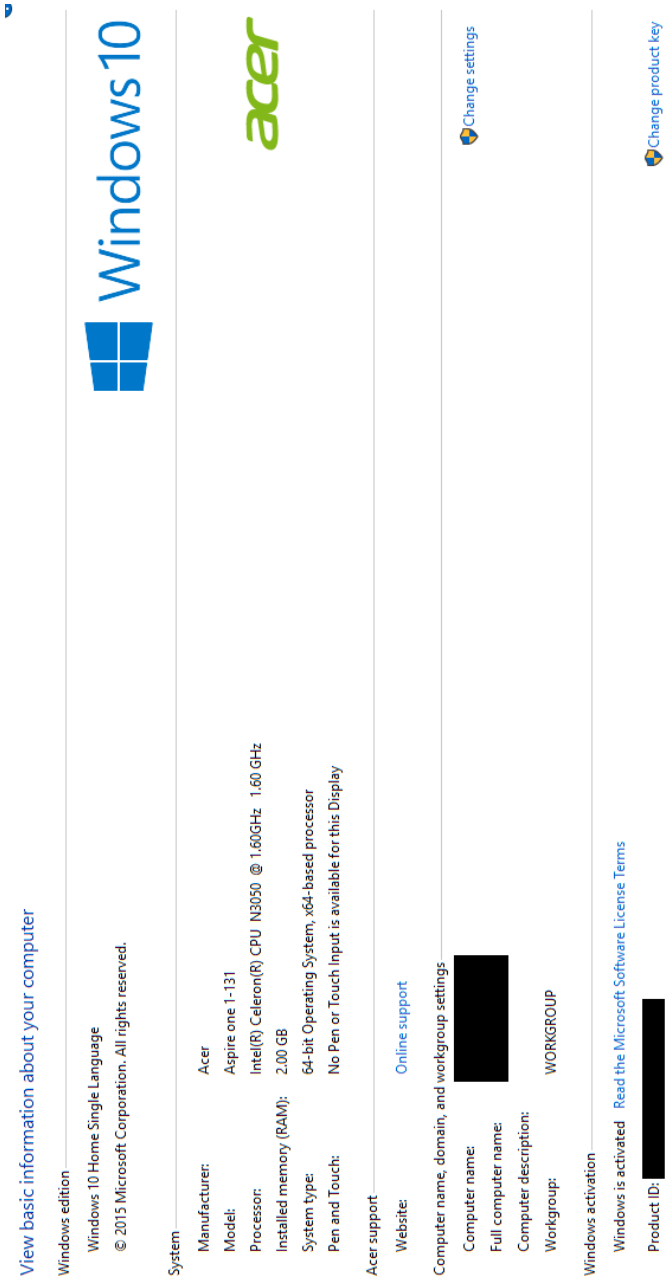
thresholding_global_otsu2.py

```
""" Python3: Global Thresholding (Otsu's Method) using OpenCV """
"""
Finds a value of 'threshold' which lies between two peaks such that
variances to both classes are minimum.
"""
import numpy as np
from matplotlib import pyplot as plt
import cv2

in_img = 'C:\\Users\\CRT13\\Desktop\\Image
Processing\\OpenCV\\images\\'+input('Choose image to process: ')
img = cv2.imread(in_img,0)
img_blur = cv2.GaussianBlur(img,(5,5),0)
# Find: Normalize-Histogram & its Cumulative-Distribution-Function
histogram = cv2.calcHist([img_blur],[0],None,[256],[0,256])
histogram_normalized = histogram.ravel()/histogram.max()
k = histogram_normalized.cumsum()

bins = np.arange(256)
fn_min = np.inf
threshold = -1

for i in range(1,256):
    p1,p2 = np.hsplit(histogram_normalized,[i])       # Probabilities
    k1,k2 = k[i],k[255]-k[i]                          # Cumulative Sum
of Classes
    w1,w2 = np.hsplit(bins,[i])                       # Weights
    # Find: Means & Variances
    m1,m2 = np.sum(p1*w1)/k1,np.sum(p2*w2)/k2
    v1,v2 = np.sum(((w1-m1)**2)*p1)/k1,np.sum(((w2-m2)**2)*p2)/k2
    # Find: Minimization Function
    fn = v1*k1 + v2*k2
    if fn < fn_min:
        fn_min = fn
        threshold = i*3
r,t = cv2.threshold(img_blur,0,255,cv2.THRESH_BINARY+cv2.THRESH_OTSU)
print('\n===Threshold===\n',t,'\n===retval===\n',r)
```

Results

Basic Dataset

Our basic dataset consisted of 5 images, captured using a *Lenovo K6 Power* smartphone. The testing system was a low-end laptop running Windows, with the following specs:



Simple Thresholding using OpenCV

OpenCV offers 5 options for "Simple Thresholding". These are listed below:

- **THRESH_BINARY**

$$\mathtt{dst}(x,y) = \begin{cases} \mathtt{maxval} & \text{if } \mathtt{src}(x,y) > \mathtt{thresh} \\ 0 & \text{otherwise} \end{cases}$$

- **THRESH_BINARY_INV**

$$\mathtt{dst}(x,y) = \begin{cases} 0 & \text{if } \mathtt{src}(x,y) > \mathtt{thresh} \\ \mathtt{maxval} & \text{otherwise} \end{cases}$$

- **THRESH_TRUNC**

$$\mathtt{dst}(x,y) = \begin{cases} \mathtt{threshold} & \text{if } \mathtt{src}(x,y) > \mathtt{thresh} \\ \mathtt{src}(x,y) & \text{otherwise} \end{cases}$$

- **THRESH_TOZERO**

$$\mathtt{dst}(x,y) = \begin{cases} \mathtt{src}(x,y) & \text{if } \mathtt{src}(x,y) > \mathtt{thresh} \\ 0 & \text{otherwise} \end{cases}$$

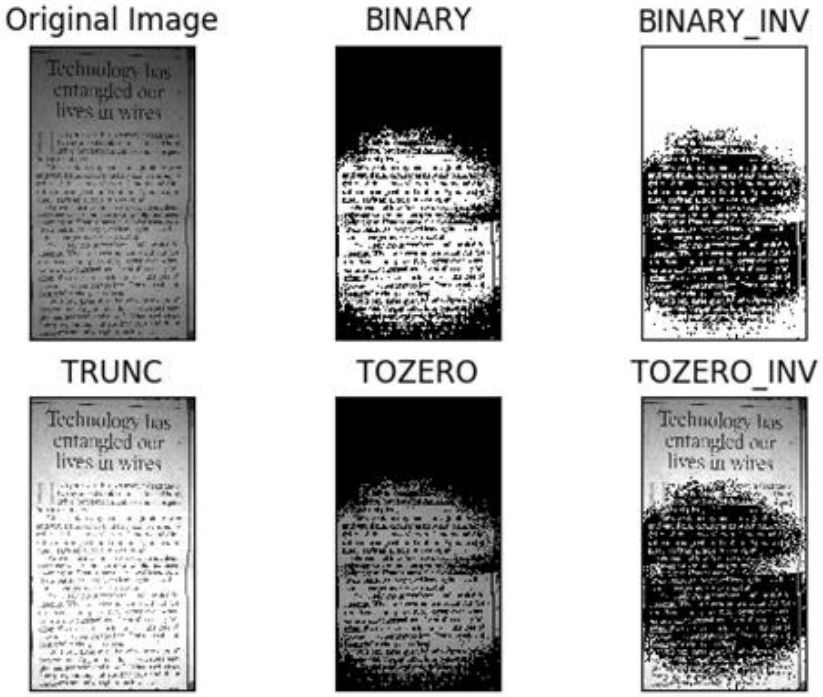
- **THRESH_TOZERO_INV**

$$\mathtt{dst}(x,y) = \begin{cases} 0 & \text{if } \mathtt{src}(x,y) > \mathtt{thresh} \\ \mathtt{src}(x,y) & \text{otherwise} \end{cases}$$

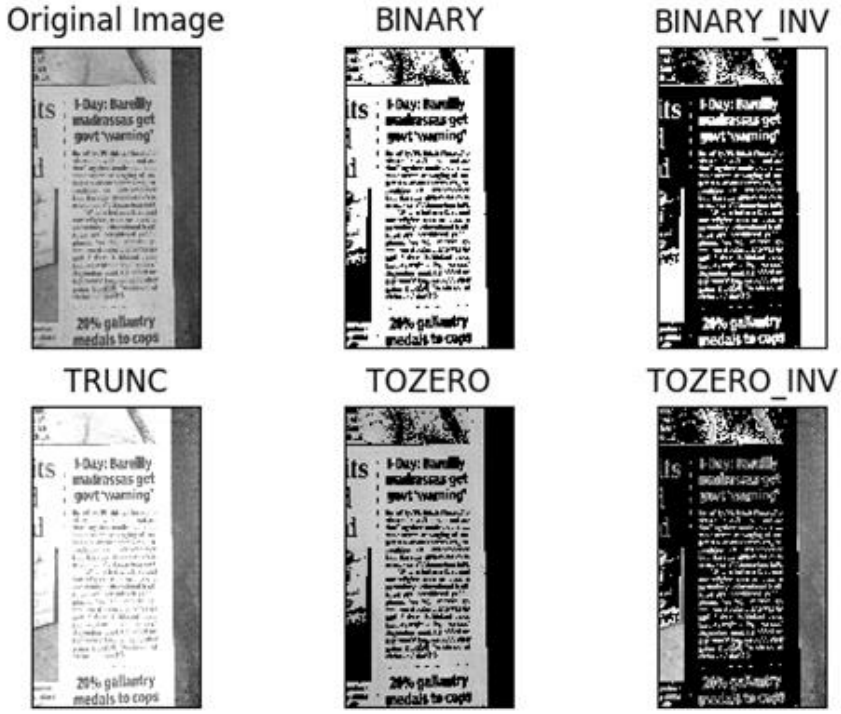
The details can be referred from pg. 294 of "The OpenCV Reference Manual". So, we generated sample plots for each of these.

THRESH_BINARY_1

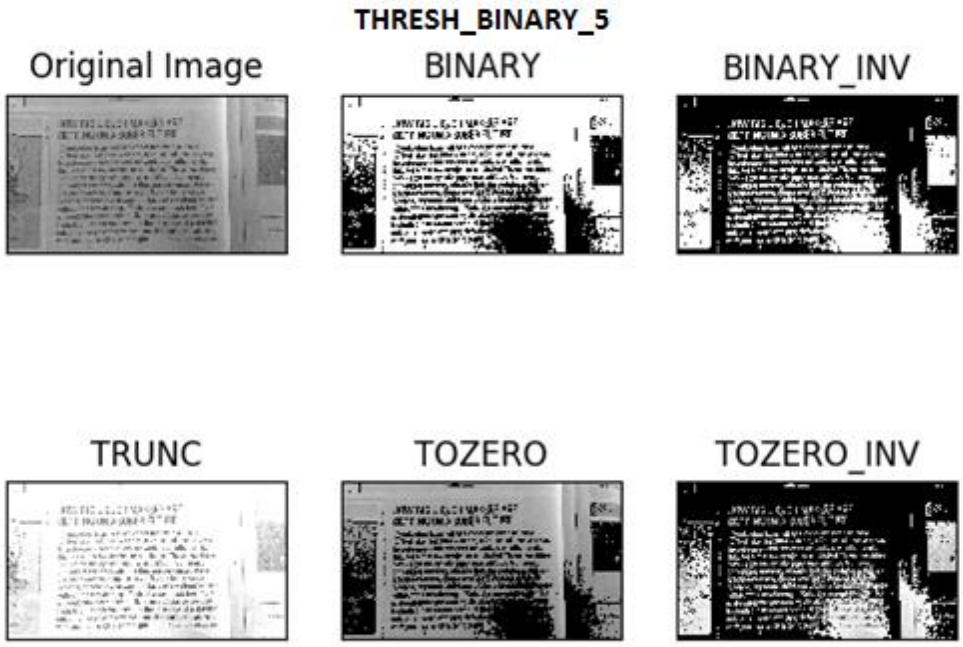
THRESH_BINARY_2



THRESH_BINARY_3

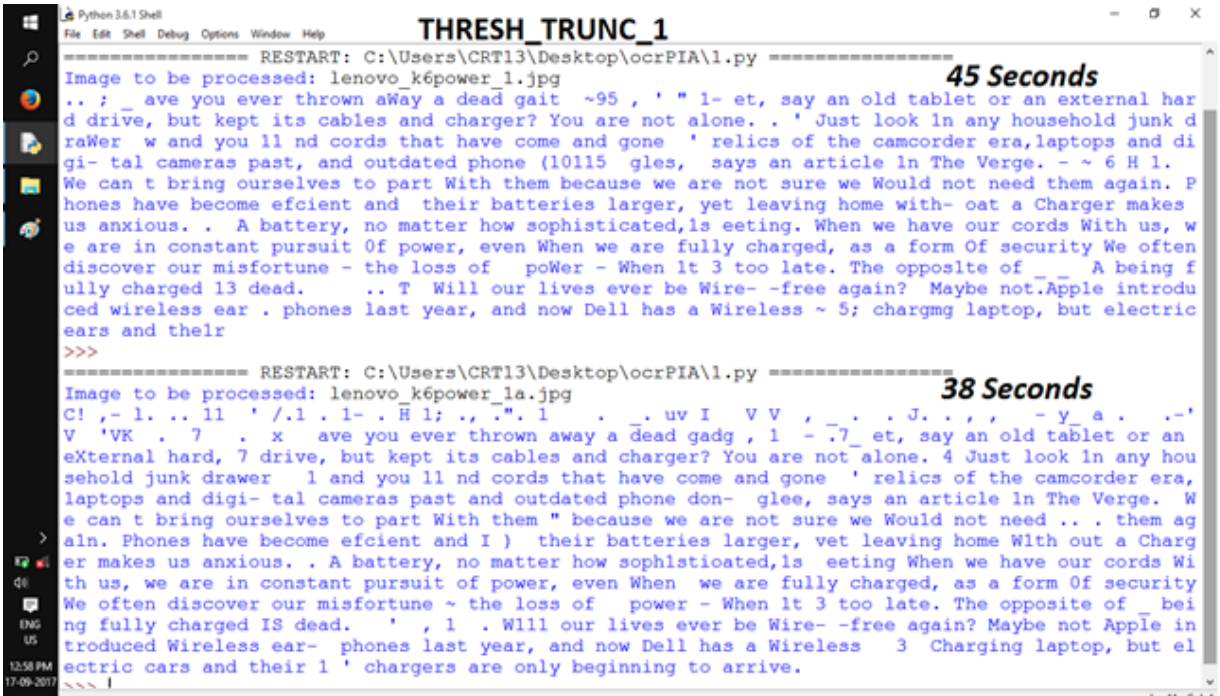


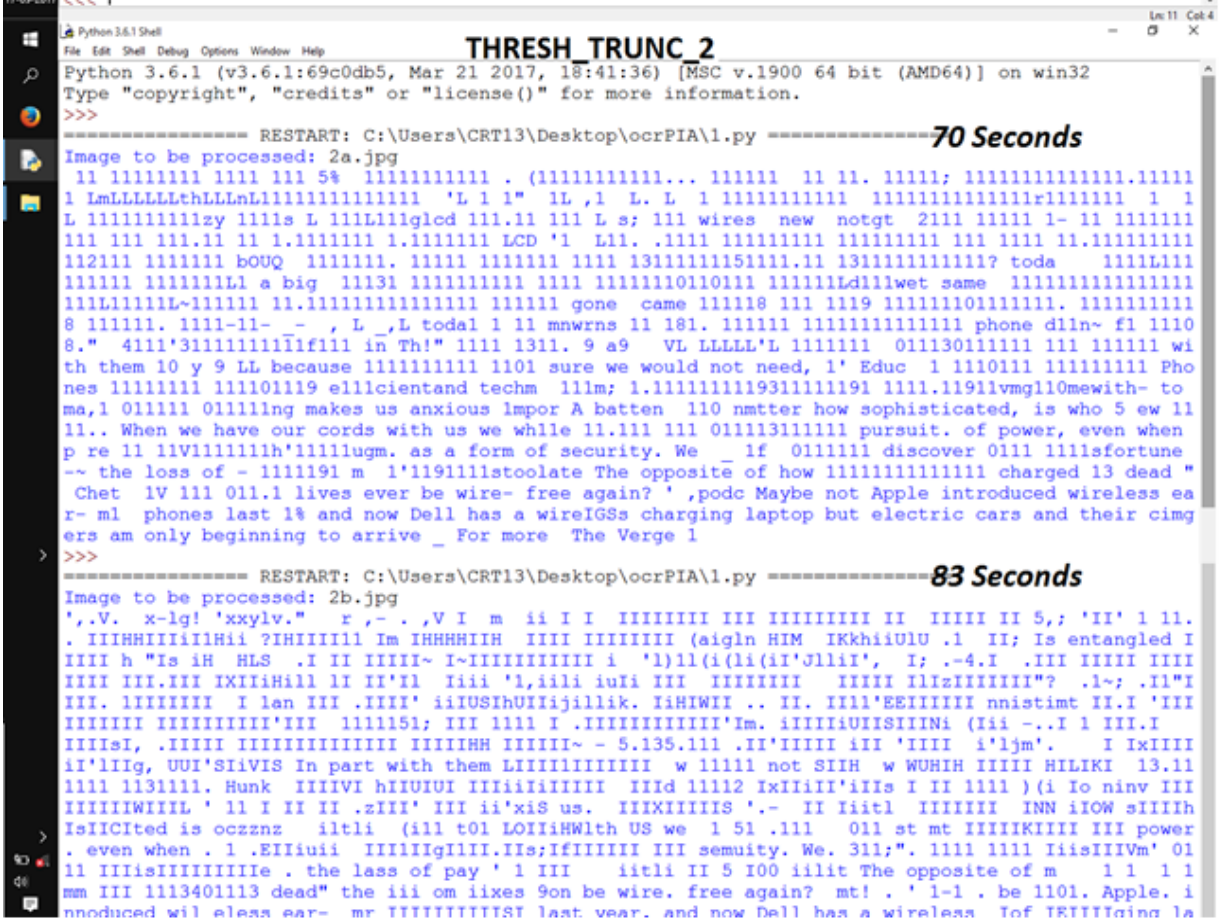
THRESH_BINARY_4



Next, we tried to observe the impact of “Simple Thresholding” on Tesseract's OCR output.

We present the results for **THRESH_TRUNC**. First things first, there was an obvious reduction in file-size, but the runtimes were arbitrary, possibly due to incosistencies caused by the lossy nature of JPEG images.


THRESH_TRUNC_1
45 Seconds
38 Seconds


THRESH_TRUNC_2
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    ml (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 ml 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 31111111111111 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 01111111141 3 1111121311113111   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        """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 lc121ry.. mlucxnkmal insti U " lm
m arr  ( rimsiderml public ' '  plums. W  will enfnrm g0~ - ; wrmmm rnder m nwdrzgxssas i 8,1").d
if  ham is 1 )laturlt 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

THRESH_TRUNC_5

```
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: 5a.jpg
(IIBIIOIII Jam; 08 8W! JHOK th - -. 53 95mm IIIIM paAomaJ .Io Ino -    pup Iaaq Ieonu 52  0 531215
u  ., Ag! Ham 2 I0 snags uImonquI aIII momIM 1.  11 ~  _, 2 I0 am am mm cum aIdoad .Io suosea; san
BIIaJ mI ) UIJIP I.uo.p OIIM a_Idoad uawo-M 102qu  SIaAIJp pamumsp am Amp JeIIIaIIM awII am m quup
 A  were 10 )IIIIJp mop oIIM aIdoad Io aAa aI'II uowo "QM s IuIIp mau 35am, IouooIe- -ou J0 IouoaI
e-MOI .IaIIIIa 3J2 mu; s IuIIp JaIIIo uo'uIzIIIOM am SIaIIIIst pm;I SJaManOJO-Iw  Aw quI Iaaq quoo
Ie ..%o 0,, I2 IIIIAA paAAoIIoI uaIIamaI-I 3123A IseI Jaaq aaII-IouooIe ue pauouneI .IasIaMpnI I
3801M 83808 V NO SNLLLBQ 38V SHBMVWHOHUII 918 MOH  ' II? 10 ill? 12 Iaaq .Iaaq quooIepu, sI  IIOII
VZIII'IUAS LAN  mu um mmmm
>>>
================= RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: 5b.jpg
Vllrnl  HOW BIG LIQUOR'MAKERS ARE BETTING ON A SOBER FUTURE  udweiser launched an aicohoI-free be
er last year; Heineken followed with a 0.0% alcohol beer this May.  Microbrewers and distillers ar
e working on other drinks that are either Iow-alcohoi or no-aicohol. These new drinks  will catch
the eye of people who dont drink or arent drinking at the time, whether they are designated driver
s, pregnant women, people who dont drink for religious reasons, or people who want the taste of a
weil-made beer  without the intoxicating effects of a weII-made beer. But  NV f iyi'ii.iiizei?irsi
x     i ,j , is no-aicohoi beer beer at all, or an artificial concoction?   ' y, . it starts off a
s normal beer and then the alcohol is distilled out, or removed with reverse osmosis (donttry maki
ng it m . "  with your home R0 filter though). , For mm  VinePojinEater  ,ugmy V " 1, ' a  v A
     5% Value
```

66 seconds

30 seconds

References

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

Literature:
[7] Eugene Borovikov, A survey of modern optical character recognition techniques
https://arxiv.org/abs/1412.4183
[8] M Seeger, C Dance, Binarising camera images for OCR (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)
http://ieeexplore.ieee.org/document/953754/
[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)
https://dl.acm.org/citation.cfm?id=1304846
[10] Ray Smith, *An Overview of the Tesseract OCR Engine* (MOCR '09 Proceedings of the International Workshop on Multilingual OCR)
https://dl.acm.org/citation.cfm?id=1577809
[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)
https://dl.acm.org/citation.cfm?id=1577804
[12] Zheng Zhang, CL Tan, Binarizing document image using coplanar prefilter (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)
http://ieeexplore.ieee.org/document/953750/
[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)
http://ieeexplore.ieee.org/document/1227732/
[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)
http://ieeexplore.ieee.org/document/953826/

Appendix

1. **Periodic Progress Report (PPR): 5**
2. **Patent Search & Analysis Report (PSAR): 5**