ICWSTCSC: 2018

TEQIP-III Sponsored

2nd International Conference

Women in Science & Technology: Creating Sustainable Career

June 28-30,2018

248: A Novel Approach of Tesseract-OCR Usage for Newspaper Article Images

Presentation by:

Chaitanya Tejaswi

Dr. Bhargav Goradiya

Prof. Ripal Patel

BIRLA VISHVAKARMA MAHAVIDYALAYA

VALLABH VIDYANAGAR

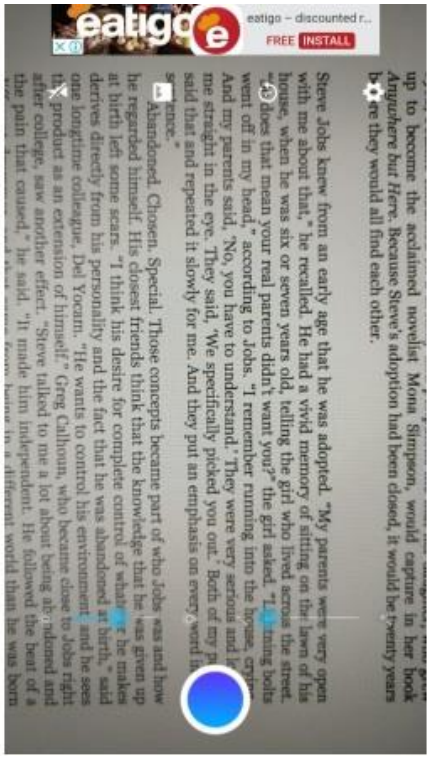
WORK IS WORSHIP

Outline of the Presentation

- Introduction
  - Scope
  - OCR: Steps
- Objective
- Tests
- Proposed Scheme
- Data Sets
- Results
- Example
- Limitations
- Conclusion & Future Scope

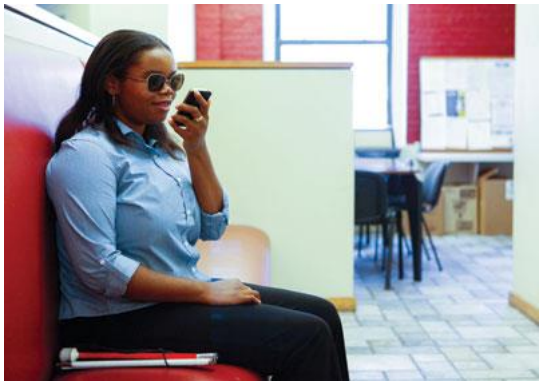
Introduction

- Our original goal was to create an Android app that gives a text to speech output for text extracted from printed newspaper articles, using open-source software tools.
- The intent was to achieve this for English language & extend it to Indian languages – (Hindi, Gujarati, …).



Scope

- *People with visual impairment*



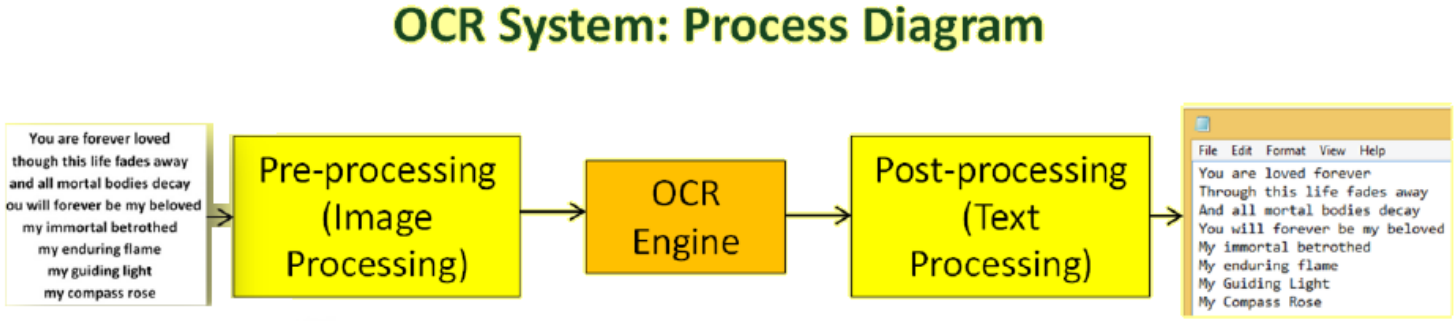
- *Journalists, who intend to deal with specific stories*

- *Daily commuters, who wish to consume daily news in an audio form*

- *Students, who wish to compile notes from a variety of sources*

ICWSTCSC - 2018

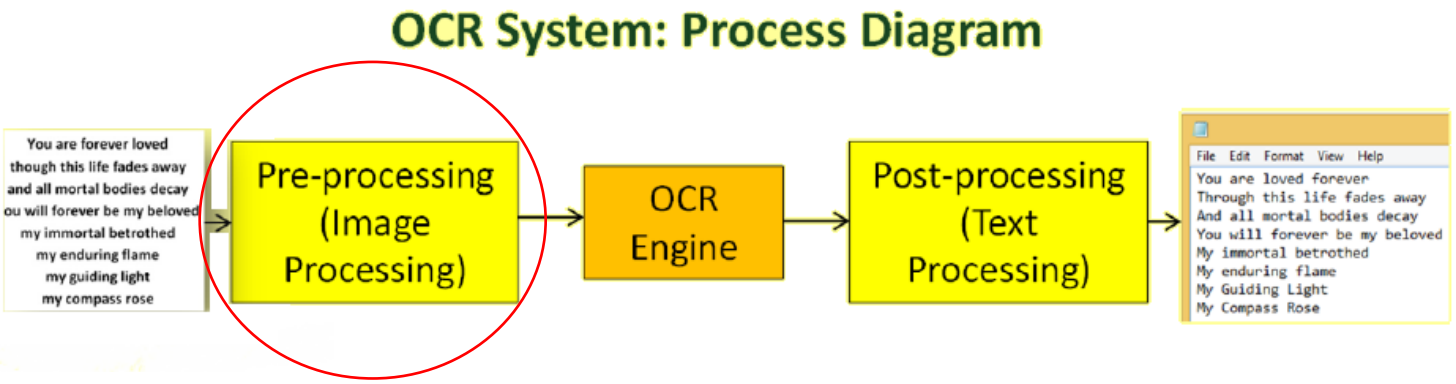
OCR: Steps



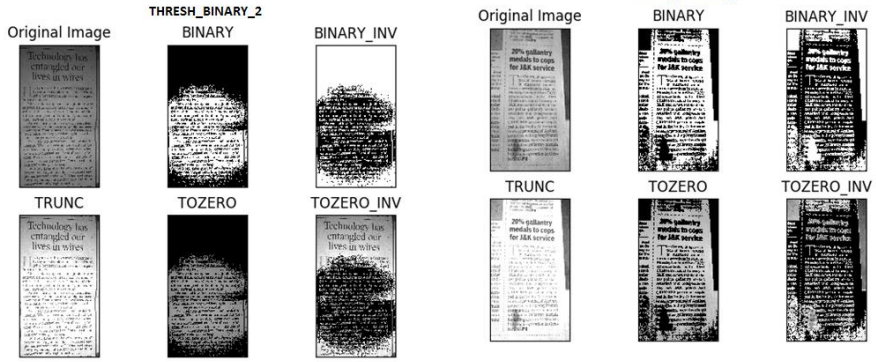
Objective

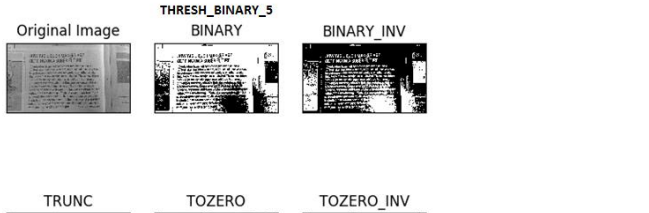
- Over time, we shifted our attention to the *Pre-Processing* stage.

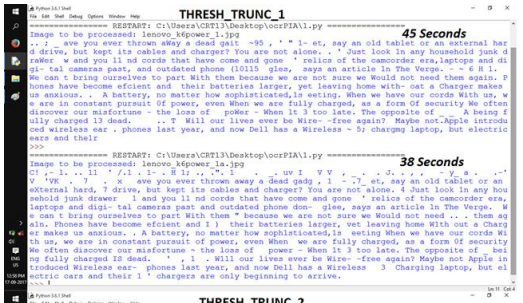
- Optimising OCR output from OCR Engine (Tesseract-OCR), by using suitable Pre-processing techniques.

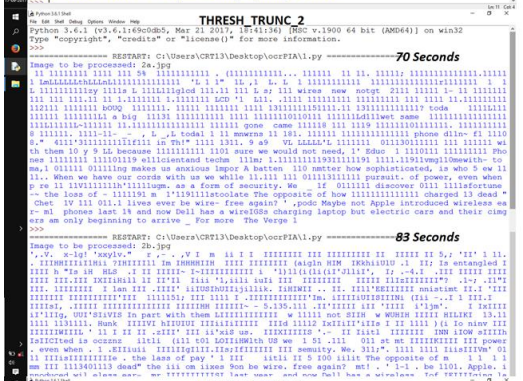


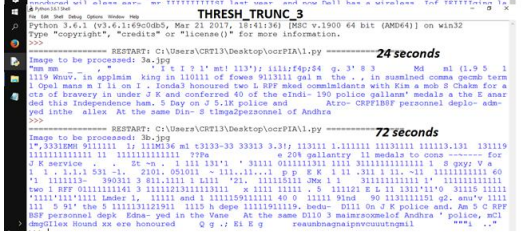

THRESH_BINARY_1
Original Image
BINARY
BINARY_INV
TRUNC
TOZERO
TOZERO_INV
THRESH_BINARY_3
Original Image
BINARY
BINARY_INV
TRUNC
TOZERO
TOZERO_INV

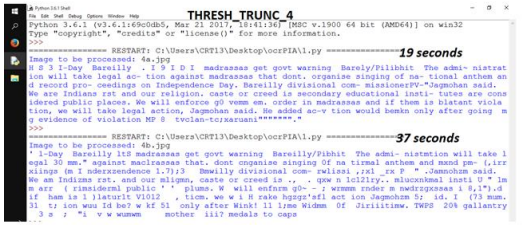

THRESH_BINARY_2
Original Image
BINARY
BINARY_INV
TRUNC
TOZERO
TOZERO_INV
THRESH_BINARY_4
Original Image
BINARY
BINARY_INV
TRUNC
TOZERO
TOZERO_INV

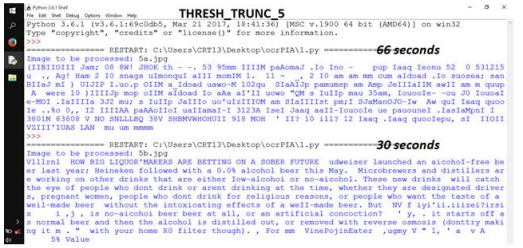

THRESH_BINARY_5
Original Image
BINARY
BINARY_INV
TRUNC
TOZERO
TOZERO_INV


THRESH_TRUNC_1
45 Seconds
38 Seconds


THRESH_TRUNC_2
70 Seconds
83 Seconds


THRESH_TRUNC_3
24 seconds
72 seconds


THRESH_TRUNC_4
19 seconds
37 seconds


THRESH_TRUNC_5
66 seconds
30 seconds


ICWSTCSC - 2018

Tests

- Determining the optimal OpenCV binarization scheme

- **THRESH_BINARY**

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

- **THRESH_BINARY_INV**

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

- **THRESH_TRUNC**

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

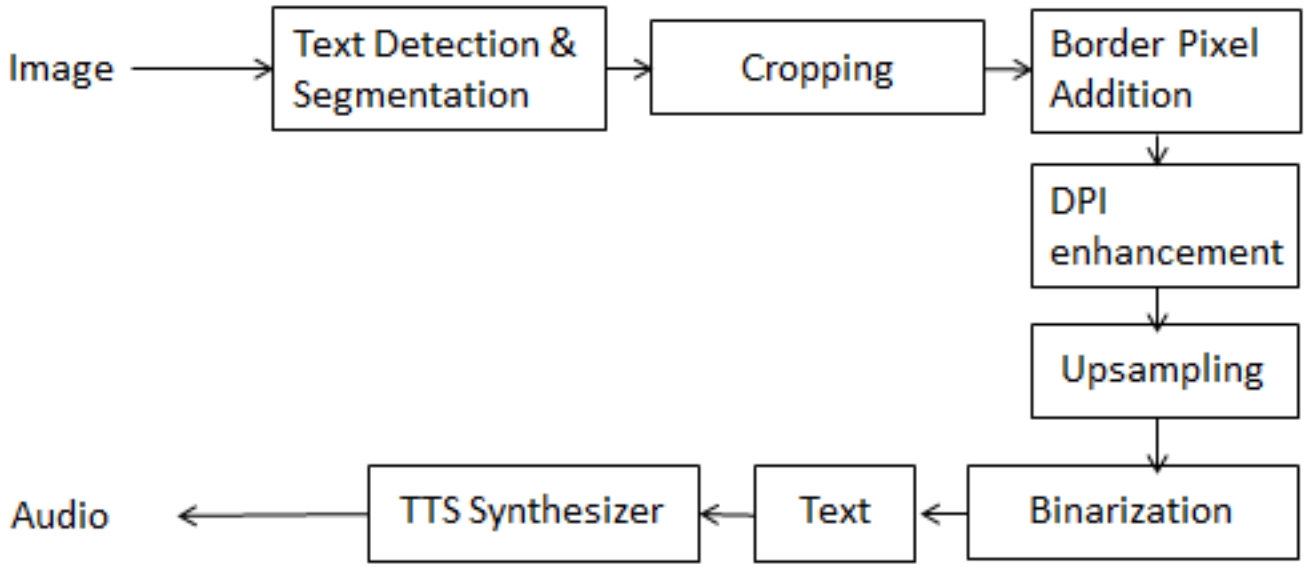
- **THRESH_TOZERO**

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

- **THRESH_TOZERO_INV**

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

Proposed Scheme



*

- *The method proposed by Lukás Neumann & Jiri Matas [16] is used for Scene Text Detection. Implemented as in-built function (OpenCV-contrib).*
- *Cropping is done by implementing Maximal Rectangle Area algorithm in Python.*
- *Upsampling is done using Gaussian Pyramid Upsampling.*
- *Binarization is done using Otsu's Thresholding.*
- *Text Synthesis is done using gTTS command line tool.*

Data Sets

1st

2nd

ICWSTCSC - 2018

Results

		Without Preprocessing		With Preprocessing		
	No. of Characters	Precision	Recall	Precision	Recall	
Set 1	3809	14.33	22.47	56.27	48.24	Worst Case
Set 2A	3620	36.2	43.33	94.4	82.13	Best Case
Set 2B	3513	44.28	48.27	69.86	78.37	
Set 2C	3508	13.33	37.81	69.86	74.25	
Set 2D	3318	15.75	32.47	64.24	68.92	

Three repetitions were made for each image, and OCR outputs were evaluated against *Ground Truth*.

Observations:

- In extreme cases, Precision was higher than Recall value. Vice-versa in general.
- Precision is usually lower than Recall values. This is due to garbage character output.

*

$$precision = \frac{number\ of\ correct\ items}{number\ of\ items\ in\ OCR\ output}$$

$$recall = \frac{number\ of\ correct\ items}{number\ of\ items\ in\ ground\ truth}$$

Example 1

File Edit Selection Find View Goto Tools Project Preferences Help

FOLDERS

02

CT01_results_original.txt

```
===== ePaper =====
===== Photos =====
===== 1 =====
[A]
P01 ' t ' Of Empathy
> In his farewel I speech, President M ukherjee made
an eloquent case for pluralism and education
11 his considered farewell speech President Pranab Mukherjee,
Ii‘ightly praised by finance minister Arun J aitley for the utterly ‘
non—partisan manner in which he has conducted himself
in Rashtrapati Bhavan, brought up the important issue of the
relationship betweenviolence and contemporary public discourse,
where he invoked the spirit of Mahatma Gandhi to argue that
when such discourse becomes tinged with intolerance, it makes
the atmosphere conducive to greater violence as well as marginaliv
sation of less privileged groups. In today’s atmosphere of polarised
debate the departing president’s fervent plea — “we must free our
public discourse from all forms of violence, physical as well as
verbal" —needs to be heard,
Pluralism has been a hallmark of Indian civilisation, and it’s
something to which we must hold fast.
,, y f, Thebattleof ideasneednotgetpersonal
" 7 ’ ' —- there are other, superior ways of
y winning it than to suggest that one's
, , opponent is anti-nationali Debate has
iii, , to be civil and rational, not hysterical
“1‘1“ .‘1 _ andparanoid;thosewho suggesttheir
“" , “‘3',” political opponents should be exiled
" ‘T “ to Pakistan are in fact making India a
‘ ' ' bit like Pakistan, which is unable to
tolerate differences. Political difference does not mean all
compassion and empathy has to be suspended — this is a truism that
1 applies equally to right and left of the political spectrum.
1 For the same reason, condemnations of violence should never be
' selective. BJPleaders’ condemnations of beef lynchings, for example,
‘ tend to be qualified and accompanied by caveats — which negate their
1 value and can be an incitement to further violence. Another issue that
5 is key to the future of the republic that Mukherjee touched upon, is
education. Unfortunately this gets stepmotherly treatment in the
political discourse, despite its critical and transformative role in
```


Line 1, Column 1 Tab Size: 4 Plain Text

[A]
[HEAD]
Politics Of Empathy
[HEAD-sub]
In his farewell speech, President Mukherjee made an eloquent case for pluralism and education
[TEXT]
In his considered farewell speech President Pranab Mukherjee, rightly praised by finance minister Arun Jaitley for the utterly non-partisan manner in which he has conducted himself in Rashtrapati Bhavan, brought up the important issue of the relationship between violence and contemporary public discourse, where he invoked the spirit of Mahatma Gandhi to argue that when such a discourse becomes tinged with intolerance, it makes the atmosphere conducive to greater violence as well as marginalisation of less privileged groups. In today's atmosphere of polarised debate the departing president's fervent plea - "we must free our public discourse from all forms of violence, physical as well as verbal" - needs to be heard.

Pluralism has been a hallmark of human civilisation, and it's something to which we must hold fast. The battle of ideas need not get personal - there are other, superior ways of winning it than to suggest that one's opponent is anti-national. Debate has to be civil and rational, not hysterical and paranoid; those who suggest their political opponents should be exiled to Pakistan are in fact making India a bit like Pakistan, which is unable to tolerate differences. Political differences does not mean all compassion and empathy hast to be suspended - this is a truism that applies equally to right and left of the political spectrum.

For the same reason, condemnations of violence should never be selective. BJP leaders' condemnations of beef lynchings, for example, tend to be qualified and accompanied by caveats - which negate their value and can be an incitement to further violence. Another issue that is a key to the future of the republic that Mukherjee touched upon, is education. Unfortunately this gets stepmotherly treatment in the political discourse, despite its critical and transformative role in India being able to meet challenges of the future.

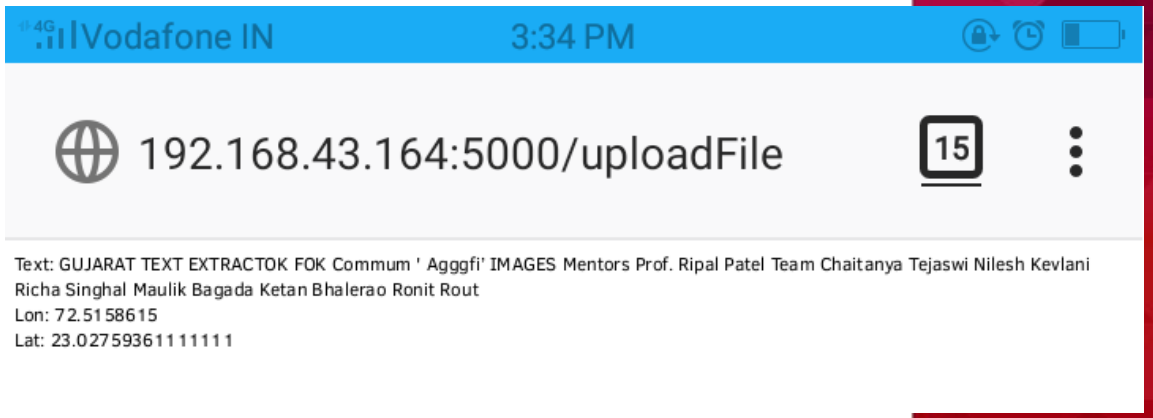
In this context, Union road transport minister Nitin Gadkari's statement that India will not allow driverless cars fearing massive loss of jobs is a giveaway of our confidence levels in the face of disruptive change. The spectre of rampant unemployment is a distinct possibility and it makes eminent sense to tame the divisive forces and violent proclivities that are in danger of cutting loose. Pluralism and education are the keys to securing India's future, and Mukherjee did well to lay down the options for both government and citizens in his expansive farewell speech.

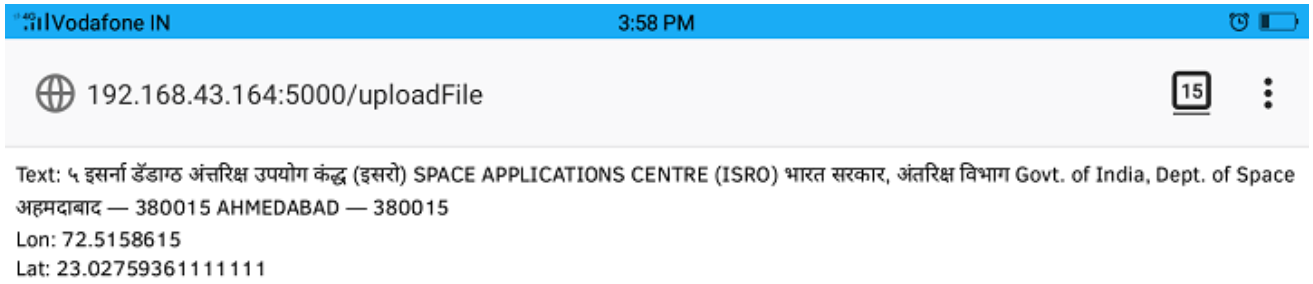
[B]
[HEAD]
Indian Cue Masters League in Ahmedabad from Aug 19
[HEAD-sub]
Solomon.Kumar @timesgroup.com
[TEXT]
Hyderabad: "For a sport that originated in the country, it is a really sad state of affairs that we have to tell people what we do. When I tell people that I am a snooker player, they seem clueless about it," India's top snooker player Vidya Pillai's lament captures the angst of the people in cue sports.

There are about 2.5 million amateur players in India but the sport still has not captured the imagination of the masses. It is considered an elitist game even though Indian players have won 42 world titles over the years. To break the ice and make the sport appeal to a larger section of the society, BSFI and Sportzlive Entertainment launched the Cue Slam Indian Cue Masters League here on

Example 2


GUJARAT
TEXT EXTRACTOR FOR COMMUNITY ASSETS' IMAGES
ARC Six
Mentors
Prof. Ripal Patel
Team
Chaitanya Tejaswi
Nilesh Kevlani
Richa Singhal
Maulik Bagada
Ketan Bhalerao
Ronit Rout


Vodafone IN
3:34 PM
192.168.43.164:5000/uploadFile
15
Text: GUJARAT TEXT EXTRACTOK FOK Commum ' Agggfi' IMAGES Mentors Prof. Ripal Patel Team Chaitanya Tejaswi Nilesh Kevlani Richa Singhal Maulik Bagada Ketan Bhalerao Ronit Rout
Lon: 72.5158615
Lat: 23.02759361111111


Vodafone IN
3:58 PM
192.168.43.164:5000/uploadFile
15
Text: ५ इसर्ना डॅडाग्ठ अंत्तरिक्ष उपयोग कंद्ध (इसरो) SPACE APPLICATIONS CENTRE (ISRO) भारत सरकार, अंतरिक्ष विभाग Govt. of India, Dept. of Space अहमदाबाद — 380015 AHMEDABAD — 380015
Lon: 72.5158615
Lat: 23.02759361111111


इसरो ISRO
अंतरिक्ष उपयोग केन्द्र (इसरो)
SPACE APPLICATIONS CENTRE (ISRO)
भारत सरकार, अंतरिक्ष विभाग
Govt. of India, Dept. of Space
अहमदाबाद - 380015
AHMEDABAD - 380015

Limitations

- **Limited to News Articles from a Single Newspaper with Consistent Fonts**


THE TIMES OF INDIA

END OF A LEGEND

Quoting Nehru, Indira & Rajiv, Modi slams Congress & Rahul

Want freedom in India, not from India: Kanhaiya

Centre rakes in big bucks from cess, but states shortchanged

Post withdrawal tax, NPS will score over EPF

Bizman's 7-year-old son abducted, killed for ₹15L

Water tables fall in 75% of state talukas

THE TIMES OF INDIA

Quoting Nehru, Indira & Rajiv

END OF A LEGEND

- **Inadequate Number of Test Images**

  10 Random, 10 Ordered

- **Angular Defects & Incident Light Intensity Variations**


Technology has entangled our lives in wires

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

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

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

"A battery, no matter how sophisticated, is fleeting. When we have our cords with us, we are in constant pursuit of power, even when we are fully charged, as a form of security. We often discover our misfortune — the loss of power — when it's too late. The opposite of being fully charged is dead."

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

Technology has entangled our lives in wires

ed a generation

Conclusion & Future Scope

- **Conclusion**

By making use of the proposed scheme on datasets of about 3500 characters each, we have observed a two-fold improvement in the character recognition efficiency of Tesseract OCR, in the case of digital newspaper article images secured using a smartphone camera.

- **Future Scope**
  - ***Extension (of model) to multiple Indian languages***

  - ***Extension to other newspaper formats***

  - ***Comparison & Objective Evaluation with other Binarization Techniques***

References

Patents:

[1] US6577762B1, Background Surface Thresholding

[2] US7400768B1, Enhanced optical recognition of digitized images through selective bit-insertion.

[3] US9298980B1, Image preprocessing for character recognition.

[4] US20120063690A1, Object-Based Optical Character Recognition Pre-Processing Algorithm.

[5] US7106905B2, Systems and methods for processing text-based electronic documents.

[6] US20130329023A1, Text recognition driven functionality.

Literature:

[7] Eugene Borovikov, A survey of modern optical character recognition techniques.

[8] M Seeger, C Dance, Binarising camera images for OCR (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition).

[9] Ranjith Unnikrishnan, Ray Smith, Combined Script and Page Orientation Estimation using the Tesseract OCR engine (ICDAR '07 Proceedings of the Ninth International Conference on Document Analysis and Recognition).

[10] Ray Smith, An Overview of the Tesseract OCR Engine (MOCR '09 Proceedings of the International Workshop on Multilingual OCR).

[11] Ray Smith, Daria Antoniva, Dar-Shyang Lee, Adapting the Tesseract Open Source OCR Engine for Multilingual OCR (MOCR '09 Proceedings of the International Workshop on Multilingual OCR).

[12] JJ Sauvola, Tapio Seppänen, Sami Haapakoski, Matti Pietikäinen, Adaptive Document Binarization (ICDAR '97 Proceedings of the 4th International Conference on Document Analysis and Recognition).

[13] Niblack, W (1986), An introduction to Digital Image Processing, Prentice-Hall, pp. 115–116.

[14] Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". IEEE Trans. Sys., Man., Cyber. 9

[15] OpenCV Reference Manual (OpenCV3.0 Documentation), pg. 265, 294-295.

[16] Lukáš Neumann and Jiří Matas. Text localization in real-world images using efficiently pruned exhaustive search. In *in Document Analysis and Recognition, 2011 International Conference on. IEEE, 2011*, pages 687–691.

Thank you