

OCR@vl2g

Abu Shahid

First Project: Automatic Grading of Indic Answer Sheets

- OCR for Indic languages
- Answer validation

- > 4k+ answer sheets from 2k+ writers
- > natural setting
 - >> lots of blur and cropped images
 - >> blank answer sheets

011

प्रश्न पत्र

• इस प्रश्न पत्र में 12 प्रश्न हैं। कृपया सभी उत्तर हिंदी में लिखें। संख्या लिखने के लिए रोमन अंक प्रणाली का प्रयोग करें (जैसे 1, 2, 3...)

• कृपया उत्तर निर्धारित बॉक्स के अंदर ही लिखें

• हम आपसे अनुरोध करते हैं कि Google जैसे खोज इंजनों का उपयोग ना करें

नोट: व्यक्तिगत जानकारी सार्वजनिक नहीं की जाएगी।
में एतद्वारा शैक्षणिक शोध उद्देश्यों के लिए अपने हस्तलिखित डेटा का उपयोग करने के लिए अपनी सहमति देता हूँ।

नाम: Anish Chauhan अनिश चौहान
आयु: 14
कक्षा: 9वीं
स्कूल/कॉलेज: रामकीर्ति उच्च माध्यमिक विद्यालय कुडकेला
हस्ताक्षर: Anish

1. x का मान ज्ञात कीजिए, जहाँ $x = 2/2 \times 2$

x का मान = 2 है

2. भारत के प्रथम राष्ट्रपति कौन थे?

भारत के प्रथम राष्ट्रपति राजेन्द्र प्रसाद थे

3. आंध्र प्रदेश की राजधानी क्या है?

आंध्र प्रदेश की राजधानी भुवनेश्वर है

4. उत्तीसगढ़ का उच्च न्यायालय कहाँ स्थित है?

उत्तीसगढ़ का उच्च न्यायालय बिलासपुर में स्थित है

5. $777 + 77 + 7 = ?$

$777 + 77 + 7 = 861$ होगा

Writer verification

> WordDetectorNN results

> DBNet results

> Craft results

श्रीमान नरेश्वर ठाकुर रामेश्वर दास मोदी

पौधों के द्वारा सूर्यप्रकाश की उपस्थिति जल व
मृदा की उपस्थिति में भोजन बनाने की
प्रक्रिया को प्रकाश संश्लेषण की क्रिया कहते हैं।

> Dataset structure for training:

<https://arxiv.org/pdf/2009.04532.pdf>

Ours??

Update- 14 january, 2023/ Auto grader @abu

- Segmentation code is ready (need samples of answer sheet)
- Augmentation done locally (need to install a library on server; [Straug](#))
- **RCNN BiLSTM CTC (no aug) Devnagri**
 - Peak accuracy: 61.5% @ 7.4e5 iters
 - 61.3% reached @ 2.4e5
 - 60.3% reached @ 22k
 - Urdu peaks at 76% @ 1.08e5 iters
 - 70% @ 24k iters
- **ResNet BiLSTM Attn (no aug) Devnagri**
 - Peak accuracy: 69.87% @22k iters
 - Resnet had more parameters as compared to RCNN
 - Training of 3e5 epochs took Resnet 1.5 times more time
 - But RCNN peaked (61%) after 5.5 hours of training
 - Resnet peaked at 70% just after 1.5 hours of training
- Detailed updates here: [Click](#)

Image name	Ground Truth	Prediction
test/14.jpg	कंपनी	कंपनी
test/29.jpg	अपयति	सपर्याप्त
test/100.jpg	ऑनलाइन	ऑनलाइन
test/350.jpg	दुष्टे	९
test/549.jpg	असफलता	शरणशरण्य
test/610.jpg	अभियान	अभियान
test/1112.jpg	फायदा	फायदा
test/1455.jpg	पेट्रोलियम	पेट्रोलियम

Few results from RCNN-CTC model

Image name	Ground Truth	Prediction
test/14.jpg	कंपनी	कंपनी
test/29.jpg	संपर्क	संपर्क
test/100.jpg	ऑनलाइन	ऑनलाइन
test/350.jpg	दुग्ध	९
test/549.jpg	असफलता	शशशशशश
test/610.jpg	अभियान	अभियान
test/1112.jpg	फायदा	फायदा
test/1455.jpg	पेट्रोलियम	पेट्रोलियम

Analysis:

Summer Challenge on Writer Verification, under NCVPRIPG'23

Contact

For any queries, please contact Abu Shahid @ shahid.3@iitj.ac.in.

Organizers:



Dr. Anand Mishra

mishra@iitj.ac.in

Assistant Professor,
Department of CSE,
IIT Jodhpur



Gyan Prabhat

prabhat.1@iitj.ac.in

Ph.D. Student



Prajwal Gatti

pgatti@iitj.ac.in

Research Assistant



Abu Shahid

shahid.3@iitj.ac.in

Undergraduate Student



Challenge:

Verify whether texts were handwritten by the same writer

आंध्र प्रदेश की राजधानी भोपाल है।

&

दिल्लीसगढ़ का उच्च न्यायालय रायपुर है।



Same Writers

आंध्र प्रदेश की राजधानी भोपाल है।

&

दिल्लीसगढ़ की राजधानी का नाम रायपुर है।



Different Writers

पंडित जवाहर लाल नेहरू

भारत के प्रथम राष्ट्रपति डॉ. राजेन्द्र प्रसाद थे

Fig 2: Sample images of Handwritten Text from different writers

भारत के प्रथम राष्ट्रपति पंडित जवाहर लाल नेहरू थे।

आंध्र प्रदेश की राजधानी भीपाल है।

Fig 1: Sample images of Handwritten Text from the same writer

> box segmentation done

> reviews needed;

>> dataset format

>> instead of NLP can we use some custom distance metrics

- What we need?: **In-house Handwritten Word Recognizer (English -> Hindi)**
Input: H/W word, output: text
- **Baselines: (i) EasyOCR and Tesseract (off-the-shelf)**
(ii) CRNN + CTC loss
(iii)

Feburary Update

- Dataset creating pipeline prepared
- STEPS:
 - 1. extract boxes from raw answer sheets and store in folders named by their writer
 - 2. split into Train-Test-Val (75:20:5)
 - 3. rename files in test and val (randomly generated 6 digit characters)
 - 4. create labels.csv (test.csv & val.csv)
 - 5. unpack Test-Val (test and val folders do not have writer subfolders; and only have images)
- Example:
- (the problem Of incomplete boxes)

```
8 > 300 images were pulled from different writers (assuming 150 writers)
9 > 150 writers were indeed extracted
0 >> 112 writer in train (8.29 images/writer)
1 >> 30 writers in test (8.5 images/writer)
2 >> 8 in val (9.5 images/writer)
3 >>> 75:20:5 ratio
4 > 1410 entities b4 unpack;
5 >> 1405/150 --> 9.366 images/writer
6 > 1372 entities aftr unpack;
7 >> 1367/150 --> 9.113 images/writer
8
9 >>> 2037 pairs in test.csv (1018 positives)
0 >>> 643 pairs in val.csv (321 positives)
1 >>> ratio of 1:1
2
```

Literature Review & Baseline Model

- Reviewed numerous papers on writer verification
- Our problem is unique
 - Text-independent
 - writer-agnostic verification
 - sparse writer data
- Implemented the [Signet paper](#) for the writer verification challenge (along with Gyan Prabhat- P21AI001)
- Achieved a performance of **0.65 AUC** on the test set, which served as the baseline for the challenge.
 - Accuracy- 63.35%
 - F1 Score- 0.72

