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। कपया सभी उत्तर हिंदी में लिखें। संख्या लिखने के लिए रोमन अंक प्रणाली का • 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 एतददवारा शैक्षणिक शोध उददेश्यों के लिए अपने हस्तिलिखित डेटा का उपयोग करने के लिए अपनी सहमति THE Anish chautem 3 1/491 - 2/514 आयुः १५ कक्षाः ७वी संकृत्वाकातेजः काध्यकीव ३५ . प्राध्यिक विद्यालय क्रेडेकेला x का मान जात कीजिए, जहाँ x = 2/2*2 ×कामल=82 % 2. भारत के प्रथम राष्ट्रपति कौन थे? अहर के जनम राख्या रावेंद्र प्रमाह वे 3 आंध्र प्रदेश की राजधानी क्या है? आंधा धेरेश की राजधानी अमरावती द

इन्तियार का उन्म नायात्र निलासपूर में मिनात है

4. छतीसगढ का उच्च न्यायलय कहाँ स्थित है?

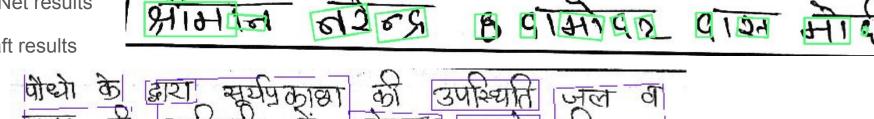
977197+92861 812

5 777+77+7 =?

Writer verification

- > WordDetectorNN results
- > DBNet results





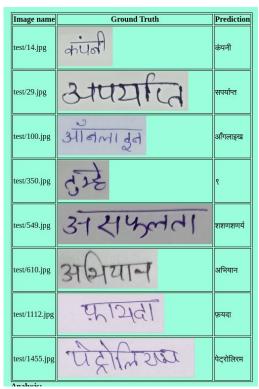
> Dataset structure for training:

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; <u>Straug</u>)
- 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: <u>Click</u>



Few results from RCNN-CTC model

Image name	Ground Truth	Prediction
test/14.jpg	कंपनी	कंपनी
test/29.jpg	अपयाति	सपर्याप्त
test/100.jpg	आवमा वैव	आँगलाइख
test/350.jpg	328	9
test/549.jpg	3-12144लता	शशणशणर्य
test/610.jpg	अभियान	अभियान
test/1112.jpg	माभ्या	फ़यदा
test/1455.jpg	ग्रेट्रीनिश्च	पेट्रोलिरम

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





Abu Shahid shahid.3@iitj.ac.in Undergraduate Student



Challenge:

Verify whether texts were handwritten by the same writer

ओंद्र प्रदेश की राजधानी भीपाल है।

&

इतीसगढ़ का उच्च त्यायलय रायपुर दै।



Same Writers

ओंद्र प्रदेश की राजधानी भीपाल है।

&

इत्तीसगढ़ की राजधानी का नाम रायपूर है।



पंडित जवाहर लाल नेहर-

भारत के प्रथम राष्ट्रपति हैं, राजेन्द्र प्रसाद थे

Fig 2: Sample images of Handwritten Text from different writers

भारत के प्रथम बाष्ट्रपति पंडित जवाहर लाल ने हरू थे।

ओंद्र 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)

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

Literature Review & Baseline Model

- Reviewed numerous papers on writer verification
- Our problem is unique
 - Text-independent
 - writer-agnostic verification
 - o sparse writer data
- Implemented the <u>Signet paper</u> 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

