# **Handwritten English Word Recognition Using a Deep Learning Based Object Detection Architecture**

Riktim Mondal

Dept. of CSE, Jadavpur University

Samir Malakar

Dept. of CSE, Asutosh College

Elisa H. Barney Smith

Dept. of Electrical and Computer Engineering, Boise State University

Ram Sarkar

Dept. of Jadavpur University

**SUMMARY:**

Due to unconstrained writing styles along with connected and overlapping characters, handwriting recognition remains a challenging task. Most of the methods in the literature use lexicon-based approaches and train their models on large datasets having near 50K word samples to achieve good results.

To this end, we propose a handwriting recognition technique to recognize handwritten English text based on a YOLOv3 object recognition model that is lexicon-free and that performs sequential character detection and identification with a low number of training samples (only 1200 word images). This is tested on the IAM dataset and it is able to achieve 29.21% Word Error Rate (WER) and 9.53% Character Error Rate (CER) without a predefined vocabulary, which is on par with the state-of-the-art lexicon-based word recognition models.

The method proposed in this research appear use similar sequential character spotting, but instead of a VGG16 Faster RCNN, YOLOv3 deep learning based object detection technique is used. This model is tested on the Latin character set using the IAM dataset.

OVERVIEW OF YOLOv3:

Object detection methods are divided into two categories based on the number of stages used to perform the task. YOLO based models only pass the input image through their single network once. The feature maps used for the detection purpose are generated following the FPN architecture. Mean square error (MSE) is used for bounding box detection and for classifying the characters, binary cross entropy (BCE) is used. The character spotting model predicts the bounding box of each character and these are compared with the ground truth bounding box to calculate the loss. The best performing model is used to recognize the test word image samples.

To assess the performance of the word recognition model, we use two popularly used performance metrics: word error rate (WER) and character error rate (CER).

WER = Number of words incorrectly recognized / Total number of words present in the test set

CER = Number of erroneously recognized characters / Total number of characters present in the test set word

We train the model using the Darknet framework that follows a transfer learning strategy starting with the YOLOv3 architecture pre-trained on the COCO dataset, and the Darknet-53 architecture pre-trained on the ImageNet dataset. Multiple bounding boxes are produced in an intermediate stage for an image, from which the best box is selected based on the object confidence score, the NMS algorithm and the maximum class confidence score.

ACCURACY:

This model is compared with several state-of-the-art methods. The two state-of-the-art models used are proposed by Sueiras et al. and Bhattacharya et al.

In the first approach, statistics is used. The goal is to determine whether there is enough evidence to “reject” a hypothesis: the proposed model does not perform well as compared to other models. The p-value is calculated, using the Wilcoxon rank-sum test from the error rates of the proposed model. If the calculated p-value < 0.05, then the null hypothesis is rejected at 5% significance level.

In the second approach, accuracy and performance scores are compared. The methods proposed by Sueiras et al. and Bhattacharya et al. obtain around 23.80% (better than the current model) and 31.30% (worse than the current model) WER respectively; their CER scores are 8.80% (better than the current model) and 13.20% (worse than the current mode) respectively. But the number of train, validation and test samples used in these works are 47952 (nearly 40 times than ours), 20306 (nearly 68 times than ours) and 7558 (nearly 0.25 times than ours) respectively. When these two models are trained and tested on our current train, validation and test datasets, the performances are poor. Our model has a 29.21% WER and 9.53% CER.

Thus, our proposed method uses significantly smaller training samples than the state-of-the-art methods while producing comparable results to them and there is no lexicon in use.

The key advantage of this model is that it performs localisation and identification of characters simultaneously in a handwritten word image using the YOLOv3 network.