## **Topic: Scanned Receipt Text Localisation & OCR**

***Proposal:*** Scanned receipts OCR and information extraction (SROIE) play critical roles in streamlining document-intensive processes and office automation in many business areas. However, commercial recipes OCR requires higher accuracy than the general OCR tasks (such as name card, license, or hand-written text recognition). Moreover, with low-quality scanned receipts and invoices, SROIE becomes more challenging in giving highly accurate results. With the aim of solving the receipt OCR challenge including text detection and recognition, this research focuses on improving localization algorithms to locate the text at different levels and develop an automatically digitized receipts OCR model where tokenizing all strings split on receipts’ space to create the ground truth - the list of words that appear in the transcriptions.

1. [**Efficient and accurate arbitrary-shaped text detection with pixel aggregation network**](http://openaccess.thecvf.com/content_ICCV_2019/html/Wang_Efficient_and_Accurate_Arbitrary-Shaped_Text_Detection_With_Pixel_Aggregation_Network_ICCV_2019_paper.html)

*Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wen- jia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel ag- gregation network. In Proceedings of the IEEE International Conference on Computer Vision, pages 8440–8449, 2019.*

The paper addresses one of the most challenging uses of convolutional neural networks in scene text detection - modelling an efficient and accurate arbitrary-shaped text detector. Pixel Aggregation Network (PAN) makes arbitrary-shaped text detectors following the simple pipeline, leading to a good balance between speed and performance, following three-fold contributions:

* Using segmentation with a light backbone network (ResNet18 or VGG16 pre-trained on ImageNet) combined with a low computation-cost segmentation head models (Feature Pyramid Enhancement Module - FPEM & Feature Fusion Module - FFM) to improve the feature representation of the network
* Proposing Pixel Aggregation (PA) to learn text similarity vector by the network and aggregate pixels nearby the text kernels
* From the predicted kernels, rebuilding complete text

Compared to other methods, while anchor-free text detectors suffer from a heavy framework/complicated pipeline slowinging down the inference speed, real time text detection is designed for quadrangular text detection which leads to the failure in locating the curve text instances.

This paper has a detailed ablation study pointing out the effectiveness of FPEM, FFM, PA as well as the influence of the number of cascaded FPEMs and the backbone to emphasize the advantages of PAN arbitrary-shaped text detector in visualization result and speed.

1. [**Efficientnet: Rethinking model scaling for convolutional neural networks**](http://proceedings.mlr.press/v97/tan19a.html)

*Tan,M. and Le,Q.V. Efficientnet:Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (ICML), 2019. https://arxiv.org/ab s/1905.11946.*

The paper introduces a new compound scaling up ConvNets method using neural architecture search balanced in all three dimensions including depth, width, and resolutions called EfficientNets to get better performance than the previous ones (MobileNets and ResNet). EfficientNets improves the ConvNet Accuracy which has challenges in hitting the hardware memory limit and ConvNet Efficiency solving the over parameterized matter, especially for super large Convests having larger design space and much more expensive turning cost.

While analysing the ConvNets scaling in one of three dimensions, this paper points out two observations :

* Scaling up any dimensions improves accuracy, but the accuracy gain diminishes for bigger models
* In order to pursue better accuracy and efficiency, it is critical to balance and coordinate all dimensions during ConvNet scaling in a principled way

Evaluated the scaling method on existing ConvNets and the new proposal method, this paper affirms that EfficientNet enable users to easily scale-up a baseline ConvNet to any target resource constraints in a more principled way, while maintaining model efficiency, surpassing state-of-the-art accuracy with an order of magnitude fewer parameters and FLOPS.

1. [**What is wrong with scene text recognition model comparisons? dataset and model analysis**](http://openaccess.thecvf.com/content_ICCV_2019/html/Baek_What_Is_Wrong_With_Scene_Text_Recognition_Model_Comparisons_Dataset_ICCV_2019_paper.html)

*Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., et al. (2019a). What is wrong with scene text recognition model comparisons? Dataset and model analysis. In Proceedings of the IEEE international conference on computer vision (pp. 4715–4723).*

The paper addresses the difficulties including the diverse text appearances and the imperfect conditions in the real world of scene text recognition (STR) models using traditional OCR methods with three main contributions:

* Analyzing all training and evaluation datasets commonly used in STR papers from synthetic datasets for training ( MJSynth (MJ), SynthText (ST)) to seven real-world benchmark datasets for evaluation mainly from ICDAR challenges previous years. Moreover, this paper reveals the inconsistency of using the STR datasets and their causes
* Introducing a unified four-stage STR framework including transformation (Trans.), feature extraction (Feat.), sequence modeling (Seq.), and prediction (Pred.) that provides a common perspective for existing methods and their possible variants toward an extensive analysis of the module-wise contribution
* Analyzing the module-wise contributions in terms of accuracy, speed, and memory demand under a unified experimental setting for both typical challenges in STR and remaining failure cases

The special things of this paper is pointing out six challenges and suggesting future research directions including calligraphic fonts, vertical texts, special characters, heavy occlusions, low resolution, and label noise. Some of those challenges appear in my chosen invoices dataset (ICDAR 2019).

1. [**A technical review on text recognition from images**](https://ieeexplore.ieee.org/abstract/document/7282362/)

*Manwatkar, P.M., Singh, K.R.: A technical review on text recognition from images. In: 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), pp. 1–5 (2015)*

This paper provides an overview about character recognition mechanisms to transform scanned documents to electronic format which also solve two challenges: first challenge is the difference between fonts of characters in paper and computer while the second one is the quality of the images when we store information from paper documents through the scanner. The architecture of text recognition contains:

* Pre-processing module: this stage tries to improve the quality of an input data (in the form of image) by noise removal, normalization, or binarization.
* Text Recognition module: with the output of the pre-processing module, this stage follows the techniques to help computers understand the character fonts including segmentation, feature extraction, and classification
* Post-processing module: the final products of this stage are in the form of text data which are understandable by the computers

This paper is not only the base knowledge for the capstone project follows but also emphasizes the wide application of text recognition technology in banking, legal, and healthcare.

1. [**A method for text localization and recognition in real-world images**](https://link.springer.com/chapter/10.1007/978-3-642-19318-7_60)

*L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In Proc. Asian Conf. on Computer Vision, pages 770–783. Springer, 2010.*

This paper focuses on solving the challenges of text localization and recognition in images with a new end-to-end method instead of using the method based on sequential pipeline processing which overall success rate is the addition of each stage:

* In recognition part, no real-world training data need to be used; method directly utilizes the characters from fonts available in the Windows OS without any preprocessing simulating acquisition effects
* Characters after the first step are performed on a representation derived from the boundaries of external regions

The proposed method of text localization and recognition avoids a pipeline architecture with a sequence of fixed decisions and introduces different novel features including: MSER detection, character and non-character classification, text line hypothesis formation, geometric normalization, character recognition, typographic model, and language model. Applying in two standard dataset, this method affirms the efficiency when improves the recognition rate from 53% to 72%, however, the performance isn’t as good as expected when using ICDAR 2003 dataset.

1. [**On the Accuracy of CRNNs for Line-Based OCR: A Multi-Parameter Evaluation**](https://arxiv.org/abs/2008.02777)

*B.Liebl and M,Burghardt. On the Accuracy of CRNNs for Line-Based OCR: A Multi-Parameter Evaluation*

The paper focuses on the tricks to train a high-quality optical character recognition (OCR) model by leveraging better network architectures and data augmentation to avoid building huge collections of ground truth. The strong point of this paper is the applicable of the method as the study approaches a wide range of OCR architectures in different tasks (modern typeface OCR, historical typeface OCR, handwritten text recognition, scene text detection). The architecture in this paper is based on Convolutional Recurrent Neural Network (CRNN) which combines CNN and RNN and needs to be careful in the trade-off between getting more details and better generalization while choosing a line height. Other than that, to get better performance of OCR, people need to find a good input data representation, find a good network architecture, find a right data augmentation, keep track of ordering the evaluation, measuring model performance, and training corpus.

1. [**Text Detection with OpenCV in Python | OCR using Tesseract (2020)**](https://www.youtube.com/watch?v=6DjFscX4I_c)

This YouTube tutorial introduces the open-source method called Tesseract to solve the OCR challenge - detecting text from images. As an open-source OCR engine, Tesseract can detect over 100 languages and use a neural network system based on Long Short-term Memory (LSTM). The drawback of this technique is the challenge in working with image having a lot of noise or the font of the language isn’t trained. Besides Tesseract, this tutorial also mentions the classic way of detecting text - using OpenCV where users can apply various manipulations such as image resizing, blurring, thresholding, or morphological operations to clean the image. After that, using OpenCV contours detection can extract chunks of data then finally apply text recognition on the contours that need to predict the text. These two open-source techniques used to be the ways I choose to tackle the ICDAR 2019 challenge. However, the accuracy rate of those methods isn’t as high as I expected.