OCR Performance Prediction for Scene Text Images with DIQA Methods: A Comparison Based Study

*Sabbir Ahmed Sibli[1], Nuray Jannat[1], Shihab Sharar[1], Annajiat Alim Rasel[1]

[1]*Department of Computer Science & Engineering, BRAC University, Dhaka, Bangladesh*

*Corresponding Author E-mail: sabbirshibli@gmail.com

ABSTRACT

Optical Character Recognition (OCR) is used to convert printed text into editable text. It is a great research field nowadays, as the study is still on the construction of perfect OCR tools. OCR is often challenging because of various barriers like low image quality, overrated text background, low text visibility, shadow problem, blurriness of texts etc. It is often difficult to extract or read the text from an image with 100% accuracy due to these reasons. Moreover, not all images are eligible for OCR purposes based on their quality. If this eligibility or OCR accuracy could be predicted earlier performing the OCR, it would save much time of performing the OCR and users would also be aware earlier if the images are not good enough. So image quality estimation is a great concern here. Again, scene text images play a great role as they are used to grab information around the environment like house number detection, car license plate number detection, billboard reading etc. As these images are captured with digital devices, focusing on the quality of scene text images is really important. In this research, a comparison has been made among some available methods to determine OCR eligibility or accuracy for a scene text image and find out which plays a better role in predicting the eligibility or accuracy. Document Image Quality Assessment (DIQA) methods that have been compared are, 1. Classical method with simple image feature, 2. Hast derivation, 3. Character gradient with sobel filter and 4. A model trained for predicting OCR accuracy directly from a given image. In this work, pytesseract has been applied as the OCR engine. "The Street View House Numbers (SVHN) Dataset" by Stanford University has been used as the dataset of this research.

Keywords: OCR, DIQA, Scene text, Character gradient, Hast derivation.