Abstract

In recent times, there have been an increase in Optical Character Recognition (OCR) solutions for recognizing the text from scanned document images and scene-texts taken with the mobile devices. Many of these solutions works very good for individual script or language. But in multilingual environment such as in India, where a document image or scene-images may contain more than one language, these individual OCRs fail significantly. Hence, in order to recognize texts in the multilingual document image or scene-image, we need to, manually, specify the script or language for each text blocks. Then, the corresponding script/language OCR is applied to recognize the inherent tasks. This is a step which is preventing us to move forward in the direction of fully-automated multi-lingual OCRs.

This thesis presents, two effective solutions to identify the scripts and language of document images and scene-texts, automatically. Even though, recognition problems for scene texts has been highly researched, the script identification problem in this area is relatively new. Hence, we present an approach which represents the scene-text images using mid-level strokes based features which are pooled from the densely computed local features. These features are then classified into languages by using an off-the-shelf classifier. This approach is efficient and require very less labeled data for script identification. The approach has been evaluated on recently introduced video script dataset (CVSI). We also introduce and benchmark a more challenging Indian Language Scene Text (ILST) dataset for evaluating the performance of our method.

For script and language identification in document we investigate the utility of Recurrent Neural Network (RNN). These problems have been attempted in the past with representations computed from the distribution of connected components or characters (e.g. texture, n-gram) from a larger segment (a paragraph or a page). We argue that, one can predict the script or language with minimal evidence (e.g. given only a word or a line) very accurately with the help of a pre-trained RNN. We propose a simple and generic solution for the task of script and language identification without any special tuning. This approach has been verified on a large corpus of more that 15.03M words across 55K documents comprising 15 scripts and languages.

The thesis aims to provide a better recognition solutions in document and scene-texts space by providing two simple, but effective solutions for script and language identification. The proposed algorithms can be used in multilingual settings, where the identification module will first identify the inherent script or language of incoming document or scene-texts before sending them to corresponding script/language recognition module.