Theme : Data Analytics

- Sub Theme : Complex Document Image Recognition using AI

There are plenty of documents which are scanned or smartphone taken in various business areas such as finance and manufacturing. In order to digitize these documents for business automation, OCR (optical character recognition) technologies have been used widely. However, in spite of their remarkable accuracies in character-level recognition, their WER (word error rate) is still too high due to the complexities of natural languages. For business automation, besides, it is desired to extract words including numbers from various types of table correctly. Unfortunately, however, it is difficult to improve the extraction accuracies using traditional machine learning methods because there exist near infinite number of table types in reality.

We are aiming to find AI based methods that extract contents of tables in an organized form from a given document image. Combined with any OCR technology, the methods have to generate spreadsheets equivalent to the contents of tables in the document image. Besides, the training of documents with various types of tables needs be done within a reasonable amount of time.

The topics we pursue through this GRO are as follows:

- Effective AI based contents extraction methods from document images including tables: separating out boxheads, stubs, and data cells with accuracy above 90%
- Lightweight model creation for machines with only CPU(s) or mobile devices without losing accuracy and speed

※ The topics are not limited to the above examples and the participants are encouraged to propose original idea.

※ Funding : Up to USD $70,000 per year