第6回中高生情報学研究コンテスト

OCR技術を用いて表をデータに自動で読み込むツールの開発

東京都立多摩科学技術高等学校　2年
佐々木哲　　柱野隆志　　小山寛史

研究背景

スケジュール管理はデジタル化が進んでいるが、未だに紙媒体を用いる機会が多くある。しかし、紙媒体では電子媒体とのデータ共有が難しく、特に電子カレンダーに予定を手入力する際、多くの労力を要する。

研究目的

学校で配布された年間行事予定表をGoogleカレンダーなどの電子カレンダーに入力するシステムの開発

学校の予定表　　読み込み　　Google カレンダーなど

開発方法

開発環境

・Google Colaboratory … Pythonの開発環境

言語

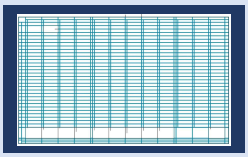
・Python (version 3.10.12)

ライブラリ

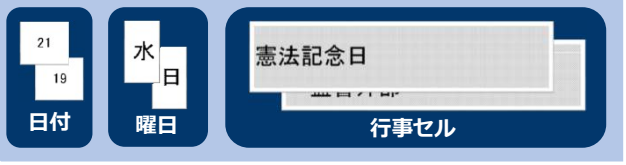
・OpenCV … 画像の二値化やセルの矩形検出
・scikit-learn … DBSCANによるデータの分類
・Tesseract-OCR … 画像内の文字を認識し文字データに変換

設計

セルの検出



セルの分類



整理したデータの出力

19	定期健康診断/学校保	2023/4/20
20	健委員会	2023/4/21
21	部活動紹介【午後】	2023/4/22

実験(プログラムの評価)

準備

・予定表の画像を入力するとGoogleカレンダーに取り込むことができるCSVファイルを生成するプログラムを作成した。

実験①

・出力されたCSVファイルをGoogleカレンダーに取り込むことができるか確かめる。

実験②

・出力された文字データの正誤を調べる。多摩科学技術高等学校の4,5月における年間行事予定と出力結果を比較する。

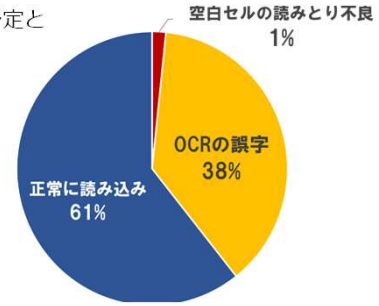
実験①の結果

出力されたCSVファイルを右図のようにGoogleカレンダーに取り込めた。

11月の予定

実験②の結果

4, 5月の年間行事予定と出力の比較結果



考察

・OCRが誤った文字をとして認識することが多い。出力された文字データを見ると以下のように誤変換されていて、文字の形状が似ていることが原因だと考えられる。

例：誤変換
「)(」→「x」　「憲法」→「意法」

今後の課題・展望

・OCRの精度向上のために画像のノイズ除去の強化する

・アンケート調査を行いどれほどの人が年間行事予定表に困っていて実際に使いやすいか調べる。

参考文献

1. 有本　寛, OCR を利用した統計表の体系的なテキストデータ化, https://hermes-ir.lib.hit-u.ac.jp/hermes/ir/re/72013/wp2021-03.pdf
2. PIAZZA, チラでジ, https://prtimes.jp/main/html/rd/p/000000110.000016981.html
3. halzo appdev, かんたんプリント管理, https://halzoblog.com/print-kanri-manual/
4. MamaLeaf,おたよりー, https://otayoly.app/