予定表の情報を電子カレンダーに追加するツールの開発

研究背景

社会全体として，スケジュール管理のデジタル化が進んでいるが学校などの場においては，予定表を紙で配布して利用している．しかし，予定を管理する上で，紙媒体と電子媒体のデータをやり取りすることは難しいと思われる．紙媒体予定を電子カレンダーで管理したいとき，その予定を手作業で入力する必要があり効率も悪い．また，このような作業を容易にするようなアプリケーションもない．すなわち，紙にある予定を自動で読み取り，電子カレンダーにその予定を入力できるツールが作業を効率化するために必要である．

研究目的

学校で配布された年間行事予定表を電子カレンダーに自動で入力することができるシステムを開発する．これによって，予定管理を電子カレンダーなどで一括で管理できるようにし，ユーザーの利便性を上げる．

事前調査

公開されている都立高校の年間行事予定表を合計で47校分を集めた．年間行事予定表には月，日付，曜日，行事などの情報が各セルに書かれている．セルの大きさは，書かれている情報によって大きさが異なる予定表が多かった．調査の結果，47校中41校の予定表で同じ要素のセルの大きさが統一されていた．

予定は曜日または日付のいずれかを基準に並べられていて，日付を基準にした予定表が47校中44校であった．

この調査より日付基準かつ各情報でセルの大きさが統一された年間行事予定表に対応したプログラムを作ることにした．

開発方法

言語

・Python (version 3．10．12)

開発環境

・Google Colaboratory … Pythonの開発環境

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テキストデータに変換

ツールの仕組み

セルの検出

予定表の画像に2値化処理を行い，2値化した画像からの表の枠線を読み取り，枠線部分のみを抽出した画像を取得する．枠線の画像から矩形検出で予定表内のセルの座標を得る．

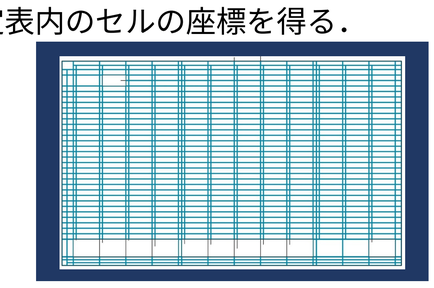

図1　年間行事予定表の直線検出

図2　年間行事予定表のセル検出

セルの分類

セルの座標情報から各セルの縦幅と横幅を基準にDBACANクラスタリングを行う．クラスタリング後，行事内容を含むセルをセルの大きさとセルの個数から推定する．

図3　セルの分類の例

整理したデータの出力

前の処理で得たセルからOCRを用いて行事内容をテキストデータ化した後，各行事と日付を結びつける．Googleカレンダーに対応した体裁にデータを作成しCSV形式のファイルを出力する．

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

図4　出力データ

実験

1．ツールから出力されたCSVファイルがカレンダーアプリに取り込むことができるか調べた．多摩科学技術高校の年間行事予定表をGoogleカレンダーに取り込んだ．
2．出力された文字データの正誤を調べた．多摩科学技術高等学校の4，5月における行事予定と出力結果を比較した．
3．多摩科学技術高等学校で4月の行事予定をGoogleカレンダーに手作業で入力した時間と，ツールを使った場合にかかった時間を測定し，ツールを評価した．

実験の結果

結果1，2

出力されたCSVファイルを図5のようにGoogleカレンダーに取り込めた．
出力結果の正誤を確認したところOCRの誤認識が多かった．

図5　Googleカレンダーのスクリーンショット

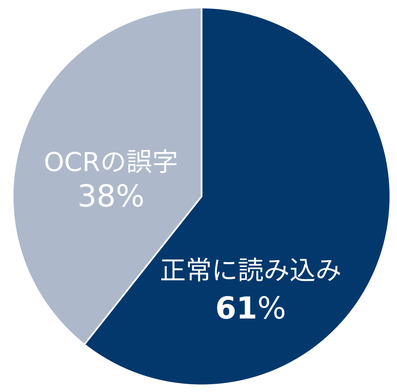

図6　出力のセルごとの出力結果の比較(4，5月分)

結果3

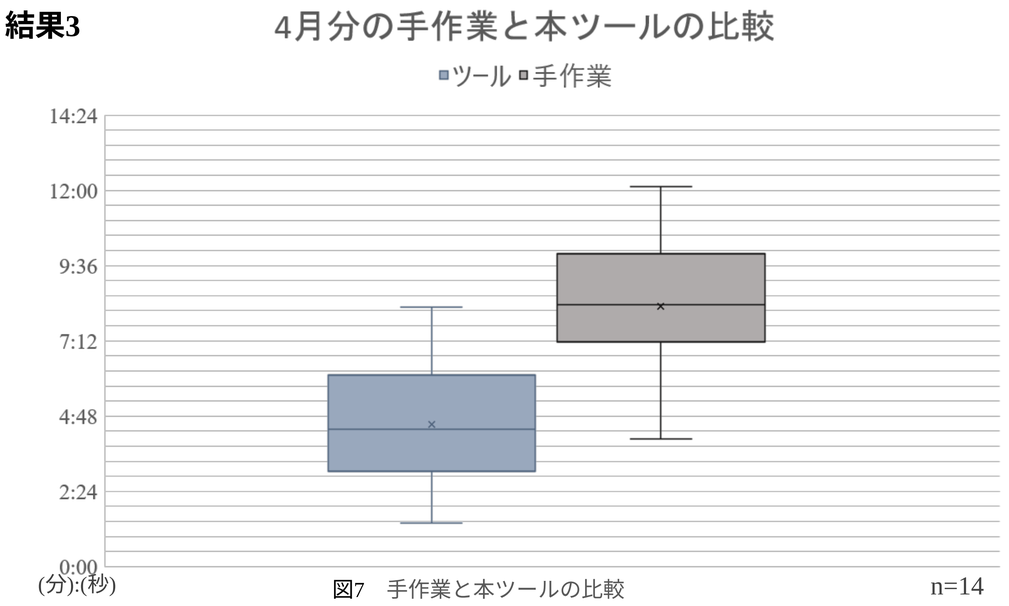

図7　手作業と本ツールの比較

4月分の入力にツールを使用して平均4分32秒，手作業では平均8分18秒かかった．

考察

・OCRが誤った文字をとして認識することが多い．出力されたテキスト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．例としては「)(」が「x」，「憲法」が「意法」になった．

・今回行った評価実験で4月分だけで平均約4分の時短効果があった．

課題・展望

・GoogleカレンダーにあるAPIをツールに実装し操作の簡略化を図る．
・対応できる予定表の数を増やす．
・OCRの誤字を減らす．

結論

・結論としては時間短縮によって年間行事予定表を扱う上でのユーザーの利便性を向上させることはできてたと言える．しかしこれは多摩科学技術高校の年間行事予定表でしか証明されていないのでさらなるプログラムの改良、調査が必要である．

参考文献

[1] OCR を利用した統計表の体系的なテキストデータ化，有本　寛(一橋大学), Center for Economic Institutions Working Paper Series No.2021-3，投稿: 2021.7.21
[2]【業界初】ChatGPTとOCRを活用した自治体DXシステムを開発，紙チラシの日付や内容を読み取り自動的にWebカレンダーで公開，情報の閲覧・管理が便利に，PIAZZA株式会社，prtimes，投稿:2023.5.12，観覧:2023.5.29，https://prtimes.jp/main/html/rd/p/000000110.000016981.html
[3]【息抜きアンケート結果発表】2023年のスケジュール管理は紙派？デジタル派？，NTTコム オンライン・マーケティング・ソリューション株式会社，投稿2023.2.10，観覧:2023.7.5 https://research.nttcoms.com/monitor/pop_info230210.html