高校生手書き数字画像データセット作成と中心位置移動による手書き数字画像の判定精度向上に関する研究

福岡県立城南高等学校　普通科理数コース ２年 小野拓登　野本悠介　１年 坂本倖嗣　永岡南見

研究の背景

近年、大学入学共通テストでの記述式の自動採点に関する研究[1]や、学校業務の負担軽減のために手書き答案の採点システムが導入されるなど、手書き文字の認識の活用が教育分野で進んでいる。

城南高校では、学校設定科目「理数ＤＳ」で、機械学習による手書き数字の認識を学んでいる。授業で利用している海外の学習用データセットでは、数字の書き方に癖があり、十分な精度が得られないことがわかった。しかし、一般に高校生を対象にした手書き数字の学習用データセットは公開されていない。また、実験を行う中で、未知画像内の数字の記入されている位置の違いにより、判定精度に違いがあることがわかった。そこで、判定精度の向上のために画像内の数字の位置を特定する必要があると考えた。

研究の目的

本研究の目的は、以下の２つである。

一つ目は、手書き数字画像の判定精度向上のため、Sikit-learnの手書き数字画像データセット（Optical Recognition of Handwritten Digits Data Set）[2]と、日本の高校生手書き数字画像データセットでそれぞれ学習済みモデルを作成し、判定精度の比較を行う。また、数字の中心位置移動により、判定精度の向上を目指す。

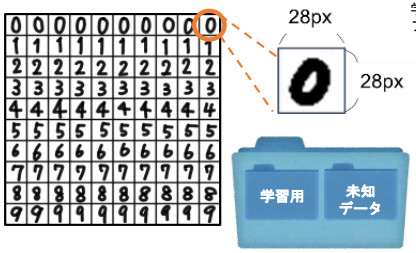
二つ目は、作成した高校生手書き数字データセットを配布し、これからデータサイエンスを学ぶ高校生や授業の中で活用してもらうことである。

高校生手書き数字画像データセットの作成

１.手書き数字データの収集方法、データセットの作成

対　象　者：高校生 計77名
使用機器：Chromebook(型番：PC-YAY11W21A4J3)
導電繊維タッチペン
収集方法：
① ペンの太さ:35、不透明度:100、ペンの種類:マーカーに設定
② 縦10マス×横10マスの方眼用紙の画像(画像サイズ2048×2048)に数字を記入
③上から0, 1, 2, 3, … , 9 の順で10種類の数字を１人10個ずつ収集(図１)
学習用データセットの概要:
①収集した画像を298×298にリサイズし、数字を28×28で切り出す
②図3のようなフォルダを作成
③数字ごとの画像フォルダに保存(図2)
収集したデータは、未知データ770枚、学習用画像6930枚(計7700枚)

図１：手書き数字のデータを収集した様子



学習用データのフォルダ構成
├dataset.0
├dataset.1
├dataset.2
├dataset.3
├dataset.4
├dataset.5
├dataset.6
├dataset.7
├dataset.8
└dataset.9

図２：データセットの概要

※未知データも同じ構成

２.学習用データセットの違いによる精度比較実験

プログラミング言語：python
ライブラリ：glob, matplotlib, numpy, pickle, Pillow
実験方法：① SVMを用いて、海外のデータセットと日本の高校生データセットをもとにモデルを作成。
② 各数字77個の未知画像を①で作成したモデルで判定。
実験結果：

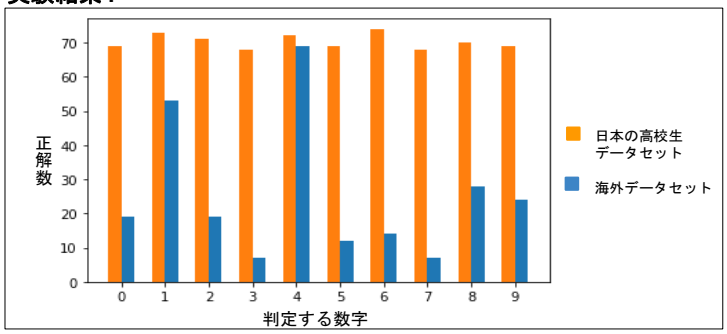


図３:正解数の比較

考察：
日本の高校生の学習用データセットで作成したモデルは正答率の平均が90%を超えた。（図３）それに対して、海外の学習用データセットは正答率の平均が約32%であった。どの数字においても、収集した学習用データセットのほうが正答率が高く、識別対象者（日本の高校生）に合わせたデータセットの作成により、判定精度が向上したといえる。

数字の中心位置の移動による判定精度の向上

１.位置による判定結果の違い

図４:位置による判定結果の違い

図４より、数字を書く位置によって判定結果[pr]が変わってしまうため、数字の位置を画像の中心に統一することで精度が向上すると考えた。

２.中心位置の特定と移動法

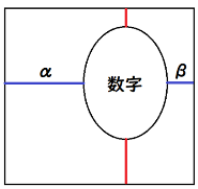


図５:数字の位置と上下左右の余白の関係

図５の青線と赤線の長さがそれぞれ等しくなるように移動する。
例）横方向
右方向へ　(β-α)/2　(※)
ずらせば良い。
縦方向でも同様に考える。

※　α + (β-α)/2 = (α+β)/2
β - (β-α)/2 = (α+β)/2
より、αとβが等しくなる。

３－１.未処理と中心位置移動の精度比較

実験①：中心位置移動（中央寄せ）が精度に与える影響を調べるために、モデルと3000個の未知画像に対して、処理を行ったものとそうでないものとで正解率の違いを調べた。
実験結果①：
処理なしの正解率は約0.29となった。
処理ありの正解率は約0.33となった。

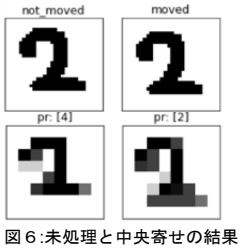


図６:未処理と中央寄せの結果

３－２．各数字の判定精度比較

実験②： モデルと未知画像それぞれに、中央寄せした場合と処理なしの場合で、0~9の各数字毎の判定結果の違いを調べるために、一種につき77個、計770個の未知画像を判定させ、正解数を調べた。結果を表１に示す。
実験結果②

表１:各数字の判定結果

モデル	画像	0	1	2	3	4	5	6	7	8	9	正解率
中央寄せ	中央寄せ	71	74	74	60	71	67	73	69	72	74	0.916
	処理なし	70	70	73	65	61	62	68	67	64	58	0.855
処理なし	中央寄せ	58	75	60	59	57	50	67	52	70	67	0.799
	処理なし	72	73	74	66	72	71	76	67	70	73	0.927

４．考察

図６や実験結果①から、画像を中心位置に移動させることが判定精度の向上に、少なからず寄与していると考えられる。

しかし、実験結果②の表１から、モデルと画像の両方に処理を行った場合、1,7,8,9の数字は処理を行わなかったものよりも正解数が多かったが、0,3,4,5,6の数字においては正解数が少なかった。これは、中央寄せ処理が後者の数字に対して、適切な処理となっていないことを示している。

これらのことから、各数字には他の数字とあまり混同せずに判定できるポイントが存在しており、前者の数字はそのポイントを中心近くに持っていると考えられる。

まとめ・今後の展望

本研究では、日本の高校生から収集した手書き数字画像データをもとに、学習用データセットを作成した。作成したデータセットと海外のデータセットで作成した学習用モデルを比較した結果、同じ数字でも、海外のデータを用いて作成した学習済みモデルより、収集した日本の高校生のデータを用いて作成した学習済みモデルの方が、正解率が高く、判定精度が向上した。今後、データの収集を継続し、データセットとして配布できるようにしたい。また、手書き数字の位置移動に関しては、中央寄せ処理では明らかな精度の向上には至らなかった。今後、各数字の最も精度が高くなる位置や特徴を特定するために、移動、回転を伴うプログラムを作成し、最も最適な位置や数字を判定できるポイントを特定することで、判定精度の向上を目指したい。

参考文献

[1] 岡知樹 ,Hung Tuan Nguyen,Cuong Tuan Nguyen,中川正樹,石岡恒憲,『大学入学共通テスト試行調査における短答式記述答案の完全自動採点』, 言語処 理学会 第28回年次大会 発表論文集P476-480,(2022年3月)
[2]『UCI Machine Learning Repository』,https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits,(参照 2023.2.5)
[3]『mgo-tec電子工作 ディープラーニングのお勉強～その11』,https://www.mgo-tec.com/blog-entry-colab-dataset01.html ,(参照 2023.2.5)
[4] https://github.com/jonan-information-contest5/information-contest5 ,(参照 2023.2.5)

謝辞

福岡県立城南高等学校情報科の棚町貴子先生からは貴重なアドバイスをいただきました。心より感謝します。また、手書き数字のデータ収集に福岡県立城南高等学校の多くの学生の皆様に協力していただきました。ありがとうございました。