JSiSE Research Report
vol.35,no.4(2020-11)

単眼カメラ画像に対する空書の自動認識

藤本　一文[*1], 長谷川 忍[*2,*1], 太田　光一[*2,*1]
*1 北陸先端科学技術大学院大学 先端科学技術研究科
*2 北陸先端科学技術大学院大学 情報社会基盤研究センター

Automatic Recognition of Kusyo by a Monocular Camera

Kazufumi Fujimoto [*1], Shinobu Hasegawa[*2, *1], Koichi Ota[*2,*1]
*1 Division of Advanced Science and Technology, JAIST
*1 Research Center for Advanced Computing Infrastructure, JAIST

The purpose of this study is to recognize Kusyo that is handwritten characters in the air automatically. We obtained the center of hand's coordinates from the color glove features from a monocular video image, visualized Kusyo by drawing the trajectory, and performed a Hiragana classification task by applying deep learning. As an experiment, we compared the four models and found good results by adding correction data and Gyousho handwriting data.

キーワード: 空書，単眼カメラ，深層学習

1. はじめに

空書とは，筆や紙を使わず，空中に文字を書くことであり，主に指 1 本で書かれる．一般に，ろう者が健聴者とのコミュニケーションを行う際には，手話が用いられるが，手話単語で表すことができないときや，文字そのものの形に意味があることを伝えたいときには空書が用いられることが多い[(1)].

空書を自動認識しようという研究としては，LED ペンを利用した研究[(2)]や, Wii リモコンやセンサを用いた研究[(3)]などが挙げられる. また, 複数のカメラを用いた手書き文字認識[(4)]や空書ではないが Kinnect を用いたジェスチャ認識[(5)]なども行われているが，十分であるとは言えない現状である.

本研究では，深層学習を利用して単眼カメラの前で書かれたひらがなによる空書を自動認識するための手法を開発することを目的とする．単眼カメラはデュアルカメラよりも安価に導入できるため環境を整えやすい．また，カメラ以外には特別なセンサを用いない非接触型の環境開発を目指す.

本研究で用いるひらがなは 50 音(ゐ，ゑを除く)，濁点，半濁点の 71 文字である．空書の検出・認識を行う上で課題として以下の 3 つが挙げられる.

(1) 空書はどこが書き出しで書き終わりなのかの識別が難しいこと.
(2) 一筆書きとなるため，線の切れ目（1 画目と 2 画目の間など）の判別が難しいこと.
(3) 深層学習を行う上で，ひらがなによる空書のデータを用意しなくてはならないこと.

本稿では，単眼カメラで収録したビデオ映像から空書を画像化し，深層学習の一手法である CNN によってひらがなの識別を行った手法および実験結果について述べる.

2. 関連研究

浅野らは LED ペンの ON/OFF を利用した空書認識に関する研究を行っている[(2)]．ペンの ON/OFF によって書き始め・書き終わりの判定や文字の切れ目の判定が容易となっている．LED ペンの輝点から方向コードを求め，輝点同士をつなぐことにより軌跡を検出している．方向コードから輝点間の距離や角度を求めることにより方向ベクトルや書く速度を求め，方向コード列の速度正規化を行うことにより個人差で変わる書くスピードの誤差を修正している．識別にはこれらのコード列データを辞書データとして登録し，未知のコー

ドと辞書データを比較する方式を採用している．

杉本らは，Wii リモコンを利用した空書による日本語入力を提案している[3]．Wii リモコンの x，y，z の 3 軸の加速度センサのうち x, z の 2 軸の加速度値を利用して筆記を検出する．また，Wii リモコンのボタンによって LED ペンと同様に書き始め・書き終わりや文字の切れ目を再現していた．また文字の再現には LSDS 法，TRRS 法を利用した．LSDS 法は文字のそれぞれの線の長さを等しい長さに分割し，それらを 8 方向コードに量子化した上で辞書パターンと照合する方法であり，TRRS 法は文字のそれぞれの線の筆記時間の割合を 8 つに量子化し，辞書パターンと入力パターンを照合する方法である．これらを用いた辞書学習を行い，文字の識別を行っている．

保呂らは，複数のカメラを用いることより，視体積交差法による人物の立体検出を行い，空書が行われた際に指先の軌跡を抽出することによって文字の検出を行っている[4]．書き始め・書き終わりに関しては，人物の立体検出により，立体形状全体の重心線から一定距離の場所に腕が出たときに判別する．文字の識別には事前に登録している文字の軌跡データとマッチングで照合し，識別を行っている．

これらの手法では，複数カメラや特別なデバイスが用いられており，手軽に意味を伝えられる空書を認識する上でのオーバヘッドが大きいことが課題である．また，近年は機械学習の研究が進み，CNN を利用して画像や文字，音声認識を行う深層学習も注目されている[6][7]．文字は傾きやサイズによって多種多様であり，日本語にはひらがなやカタカナ，漢字といった文字が多く存在する．多種多様な文字への対応として，CNN は汎用性が高いが，学習するための膨大なデータセットを用意しなくてならない．研究では公開データセットを用いることが多いが，日本語に関する空書の大規模な公開データセットは現時点では存在していない．

3. 提案手法

3.1 データ収集

空書を行う際には，対面から見ることが多いため，本研究では単眼カメラを対象者の対面に設置し，ひらがなの空書の様子を動画として取得することとした．

そこでまず，空書の動画データを作るために被験者に 50 音(ゐ，ゑを除く)，濁音・半濁音の 71 文字の空書を行ってもらい，その様子を被験者の対面に設置したビデオカメラ(Sony FDR-AX100)で 1,920×1,080 画素 30fps で撮影した．明るさや背景の条件を統一するために動画データの収集はすべて学内の 1 室で行った．また，識別の時に文字が混ざらないようにするために，1 文字ずつ順番に間隔をあけて書いてもらうようにした．動画データ収集の環境について図 1 に示す．

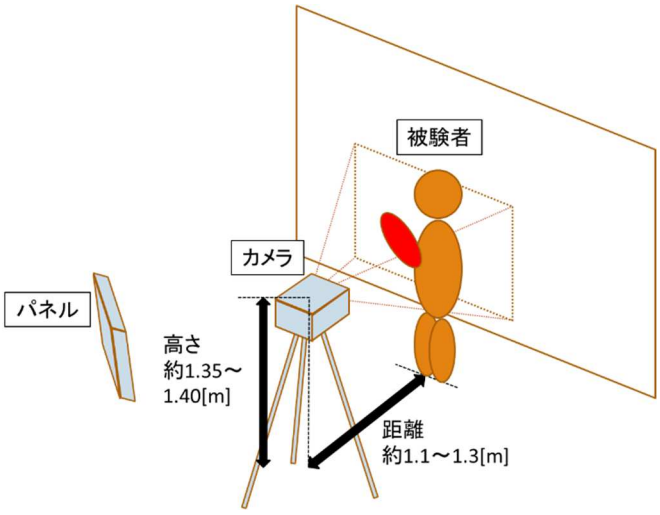

図 1. 動画データ収集の環境

空書を行う際，ビデオカメラの撮影している範囲内で行ってもらう必要がある．そこで，ビデオカメラのモニタを被験者側に向け，空書が画面内に収まっていることを確認してもらいながら収録を行った．また，被験者が 71 文字を書く上で，次にどの文字を書くべきかわからなくなることが懸念されたため，別のパネルもしくはディスプレイを用意し，そこに次に書くべき文字が表示されるようにした．

収録時，被験者はカメラに向かって文字を書いているため，動画データ上では文字が反転した状態で収録されている．しかし，動画データの読み込みの時に反転することで，文字画像は通常の向きで扱えるようにした．

3.2 線の描画

今回は，背景が白い場所で被験者の手に赤色手袋を装着し，色抽出処理を行うことにより手の領域を検出することとした．検出時には，空書の動画データから得られた RGB 画像を利用する．まず，RGB 画像を HSV

画像へと変換することで,明るさの変化に頑健にする．次に,予備実験を通じて設定した赤色の HSV の閾値により手領域の抽出を行う．図 2 に，赤色手袋が検出されている様子を示す．

図 2 赤色手袋の検出

図 3 に書きはじめ・空書途中・書き終わりの流れを示す．本手法では，領域の特徴量が動画の範囲内で一定以上となるとき，文字の書きはじめと判断する．逆に一定以下となったときを書き終わりと設定する．ただし，このままでは特徴量に左右されやすいため非常に不安定である．そこで，特徴量が一定以下となってから 0.3 秒ほど時間を置き，その間に入力が行われなかったと判定された場合に書き終わりと判定する．これにより，誤検出や文字を書いている最中に文字が分裂することを低減する．また，動画内に文字を書く予定のない手の動きが混入すると，画像生成時にノイズとなり，文字認識に支障をきたす恐れがあるため，画像を生成する段階で，画像サイズが条件に満たないものは除外されるようにした．また，空書途中に被験者の手が画面外に出てしまったものについては，文字が分裂してしまうため，学習データには加えないこととした．

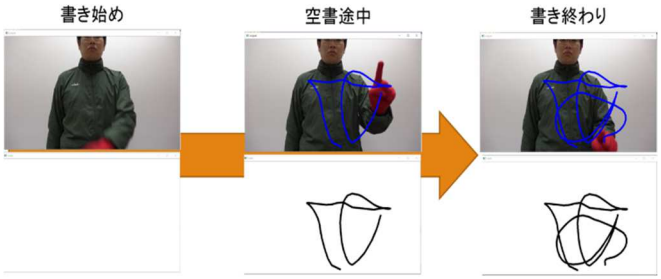

図 3. 空書で「あ」を書く際の例

3.3 線の切れ目の再現

空書を行う際，空中の文字は常に一筆書きとなってしまうため，文字の切れ目を判別することが困難である．本手法では，線の切れ目を再現するために，以下の条件を設けた．ここで対象としているひらがなは基本的に上から下に向かって書いていく．ひらがなを図 4 にある一つのマス目に書く時，色のついている左上の領域内から書きはじめることがほとんどである．そしてそこから線を引き，文字を書いていくが，その最中に左上の領域に戻ることはほとんどない．そのため文字を書いている際に左上に向かう線を削除対象とすることで，線の切れ目を再現する．

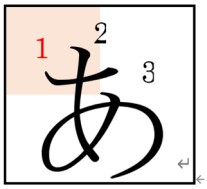

図 4. ひらがなの書き位置

線の軌跡は抽出された手領域の重心の座標の移動であらわされる．方向ベクトルを図 5 に示す．左上に向かう線は重心の移動前の点 P_(i-1)から移動先の点 P_i の角度パラメータ θi で求めることができる．角度パラメータ θi を求める方法として式(1)に示す. θi の範囲を(2)に示す．条件として左上に向かう線であるため，求められる θi の範囲は真上に向かうもので 90°，正面向かって真左に向かうもので 180°の範囲を条件とする．

$$\theta_i = \tan^{-1}\left(\frac{y_i - y_{i-1}}{x_1 - x_{i-1}}\right) \qquad (1)$$

$$90 < \theta_i < 180 \qquad (2)$$

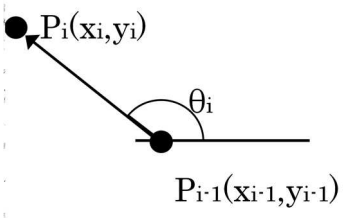

図 5. 方向ベクトル

3.4 文字認識

空書は人が直接書いているため，手書き文字に酷似していると言える．CNN による手書きの日本語認識については多くの研究[6][7]がされており，本研究では，空書と手書き文字の類似点に着目して，文字認識に CNN を用いる．

CNN のトレーニングを行うためには，大量の訓練データが必要であるが，空書の動画・画像データを十分

数用意することは困難であると考えられる. そのため, 本実験では少ない学習データでトレーニングが行える Fine Tuning を適用する．また，Fine Tuning を行うための学習モデルとして ResNet50 を用いる.

3.4.1 ResNet50

Residual Networks（ResNet）は, 2015 年の ILSVRC でトップとなった Kaiming He らのネットワークである[8]. 一般に, ある程度の多層ニューラルネットワークは層が少ないニューラルネットワークよりも精度が高くなるが，あまりに多いと勾配消失問題が発生し精度が悪化する．なぜなら，各層ごとに活性化関数の微分を行い，勾配を計算することで重みを調整しているが，層を増やしすぎると微分の積が多くなりすぎて勾配が消えていくためである．ResNet では通常のネットワークのように，何かしらの処理ブロックによる変換 F(x) を単純に次の層に渡していくのではなく，その処理ブロックへの入力 x をショートカットし, H(x)=F(x)+x を次の層に渡すことが行われている．このショートカットを含めた処理単位を residual モジュールと呼ぶ．ResNet では, ショートカットを通して, backpropagation 時に勾配が直接下層に伝わっていくことになり，非常に深いネットワークにおいても効率的に学習ができることが利点である.

3.4.2 Fine Tuning

Fine Tuning とは，深層学習において既存のモデルの一部を再利用し, 新しいモデルを構築する手法である. 前半部分は画像の特徴の一般的なことを捉えているため再学習させる必要がない．後半の層になるほど，より具体的な特徴を捉えるようになっているため，最後の層のみを再学習させることで新しいクラスを識別できるようになる．公開されている多くの汎用モデルを使うことで，特定用途のモデルを構築することができる. また, モデルを作る際には, 1 クラスあたり数千～数万の訓練データが必要になるが，Fine Tuning を用いると再学習に必要なデータは 1 クラスあたり数十～数百の訓練データだけでモデルを作ることができるため, 本研究のように用意できるデータが少ない場合に効果的である.

4. 評価実験

4.1 データセット

29 人の被験者にビデオカメラの前で 71 文字を順番に書いてもらうことを 2 回行ってもらい，1 クラスあたり約 58 枚のデータを収集した. しかしながら, 深層学習を行う上では 1 クラスあたりのデータ数が少なく, 認識精度の向上が見込まれないため，データ加工を行い，データ数を増やした．図 6 に 3.3 節の補正前の画像データ，図 7 に補正後のデータの例を示す．なお, 文字の大きさは縦横の比率で大きく見えてしまっているが，文字自体はほぼ同じ大きさである.

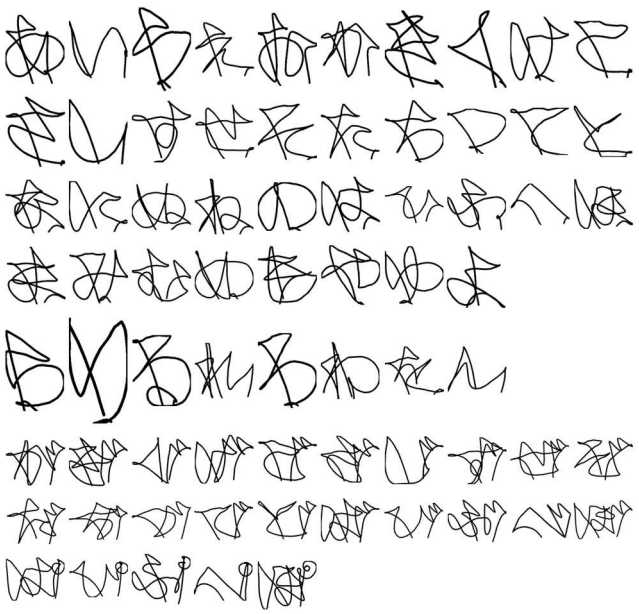

図 6. 補正なし画像データ

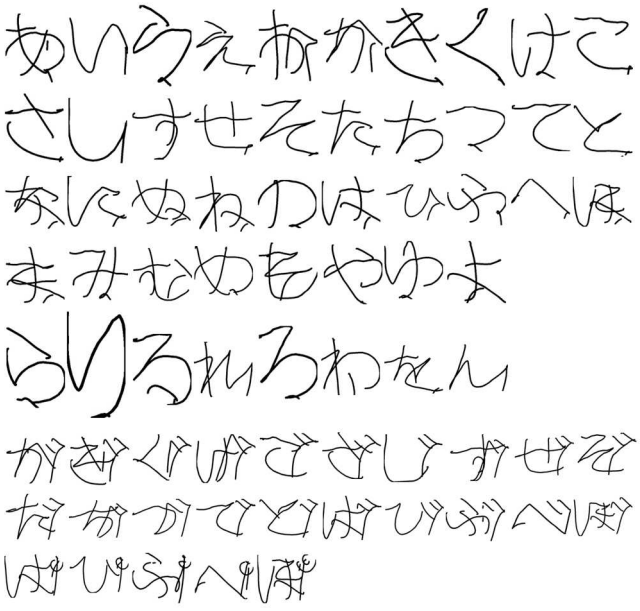

図 7. 補正あり画像データ

また，流れるように書かれることの多い草書体は文字の切れ目が繋がっていることがある．そういった草書体を含めたひらがなの手書き文字の画像データセット[9]が公開されているため，本実験では訓練データに追加したものも作成した．

データ加工では，1 枚の画像を x，y，z 軸で左右に回転したものを用意した．x 軸，z 軸のものはそのまま回転させてしまうと画像が乱れることが多かったため，本実験ではアフィン変換による疑似回転したものを使用する．これにより元画像と合わせ 5 パターンの画像を作ることができる．さらにこれらの y 軸で左右に 10°ずつ回転をさせたものを用意したため，1 クラスあたり約 870 枚まで増やすことができた．x，z 軸で疑似回転した画像を図 8 に示す．黒塗りされている部分は回転を見やすくしたものであり，学習データとして使用する際には白色になっている．

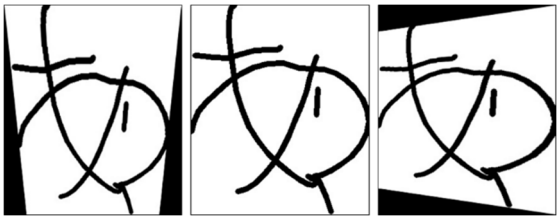

図 8. x 軸，z 軸で疑似回転させた画像例

4.2 条件

本実験では，表 1 に示す 5 つのモデルを構築してその精度について比較を行った． 1 は，線の切れ目の再現を行っていない補正なしのデータでモデルを構築したものである．2 は上に向かう線を削除した補正ありのデータでモデルを構築した．3 以降は補正ありデータに対して，手書き文字データを追加したモデルとなっている．4 は，手書き文字データで 50 音（ゐ，ゑを除く）部分をさらに加算したものとなっている．これは，50 音と濁音，半濁音で追加データにより影響があるのかの評価を行うためである．

学習モデルを作成するための学習では，学習データの量が少ないため，学習済みの ResNet50 モデルに対する Fine Tuning を行った．これらで学習する際の設定したハイパーパラメータは，batchsize＝128 ，max_epoch =40 に統一し，5 分割交差検証を行った．

表 1. 構築したモデルの条件設定(数値は枚数)

	データ	Training	Validation	Test
1	補正なし	49,116	12,279	710
2	補正あり	49,116	12,279	710
3	補正あり＋手書き(100)	54,796	13,699	710
4	補正あり＋手書き (3+50 音 100)	58,476	14,619	710

4.3 方法

構築した各モデルに対してテスト用の動画データを利用してひらがなの認識精度の評価実験を行った．評価実験では，学習データで集めた 29 人とは別の 5 人の被験者に 2 回ずつ 50 音（ゐ，ゑを除く）と濁音・半濁音のひらがな 71 文字の空書を行ってもらった. それらのデータから認識精度を比較した．被験者ごとの認識精度は 71 文字を通した識別率を求めた. 文字クラスごとの認識精度はそれぞれのクラスで 10 回行った識別のうち誤検出された回数から求めた．どちらの認識精度も求める際には(3)式で求めた.

$$\text{識別率} = \frac{\text{識別回数}}{\text{識別回数＋誤検出}} \qquad (3)$$

4.4 結果

モデルごとの 71 文字全体の識別率を表 2 に示す. なお，5 分割交差検証で 5 つのモデルを構築し，同一のテストデータでそれぞれ実験を行った．表 2 では Answer が正しく書かれた数，Miss が間違えた数，Average が識別率を表している．また，50 音と濁音・半濁音を分けた識別率も示した．これらの結果から，補正なしのモデル 1 でも，96%という識別率であり，補正を加えたり，手書き文字を追加することで，98%まで識別率を向上させることができた.

表 2. 各モデルの識別率

	モデル			
	1	2	3	4
Answer	3,434	3,468	3,482	3,483
50 音	2,264	2,275	2,276	2,282
濁音・半濁音	1,170	1,193	1,206	1,201
Miss	116	82	68	67
50 音	36	25	24	18
濁音・半濁音	80	57	44	49
Average	0.967	0.977	0.981	0.981
50 音	0.984	0.989	0.990	0.992
濁音・半濁音	0.936	0.954	0.965	0.961

4.5 考察

モデル毎に間違えた割合が多いひらがなについて，補正を行わなかった場合と行った場合とを比較し，分析を行う. 複数回の誤認が起きている文字については，同じ被験者の同じ文字が 5 分割交差検証によって生成された各モデルで同じ誤認を起こすことがしばしば見られた．そこで，ここでは各モデルの誤認の合計が 5 回以上の文字を中心に分析を行う.

4.5.1 「あ」について

「あ」のひらがなでは，実験を通して「お」,「め」に誤認を起こすケースが見られた．誤認が起きた文字の補正なしと補正ありの文字を図 9 に示す．補正がかけられていないときは,「お」と「め」の複数の誤認が起きていたが，補正をかけた後は「お」の誤認はみられなかった．上方向の線を消した効果はあったと思われる．だが,「め」の誤認は補正をかけた後にもみられた．誤認を起こした理由としては,「あ」の 1 画目と 2 画目の線が 3 画目の線と比べ短く，全体的に「め」と誤認が起きやすかったと思われる.

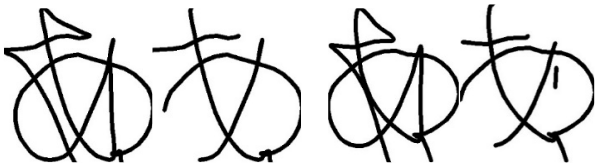

図 9. 「お」,「め」と誤認した「あ」

4.5.2 「い」について

「い」のひらがなでは，補正なしの状態の時，2 人の被験者らのひらがなが「り」と誤認された．補正をかけた後や手書き文字データを追加したモデルでは 2 人の「い」は誤認を起こすことが少なくなり，精度は良くなった．上方向の余分な線がノイズとなり，精度を落としていたと思われる．また，補正をかけた後,「へ」と誤認する比率が高くなったケースも見られた．これは，被験者の 1 人の「い」が横に長くなってしまっていたため，補正をかけた後は「へ」に見えるようになってしまったためである．誤認が起きた文字の補正なしと補正ありの文字を図 10 に示す.

図 10. 「り」,「へ」と誤認した「い」

4.5.3 濁音・半濁音について

濁音・半濁音は全体的に 50 音と比べると識別率が低かった．濁音・半濁音の 25 文字は 50 に濁点・半濁点がついているかで識別しなくてはならない．また，画数が多くなるため，手を動かす箇所が多くなり，軌跡が長くなる．そのため，余分な線が多く含まれノイズが生まれやすくなる．特に「は行」,「ば行」,「ぱ行」は濁点・半濁点の有無に加え，さらに濁点・半濁点の区別も行わなければならないため，識別が困難であった．特に「ぷ」や「ぽ」は画数が多いひらがなであり，誤認した回数が最も多かった．濁点や半濁点を書いたとき，書き方によっては点や丸の上を横断してしまっている．誤認が起きた文字の補正なしと補正ありの文字を図 11 に示す.

図 11. 「ぶ」「ぼ」と誤認した「ぷ」「ぽ」

これらをまとめると, 71 文字すべての認識精度で見たときは，補正をかけなかったものより補正を加えた場合や手書き文字データを追加した場合で認識精度は良くなっていた．しかしながら，特定のの酷似したひらがなや画数が多いひらがなだけ見たとき，書き方によって認識精度はあまり向上しなかった．また，補正を加えたモデルでも認識精度がよくないときは手書き文字データを加えたモデルも認識精度は良くなっていなかった．これはデータ加工を加えたとはいえ，元の画像の量が少なかったことから，データ不足の可能性は考えられる．また，補正を加えたことによって認識精度が落ちたひらがなもあった．これは上方向の線を消すということによって不必要な線と必要な線の判別が行えていないため，識別に必要な線まで消してしまっていると思われる．このため，不必要な線と必要な線の判別を行える方法を検討する必要がある．また，濁音・半濁音を 50 音と一緒の識別器で行うのではなく別の識別器で認識する方法も考えられる.

5. おわりに

本研究では，深層学習を利用して単眼カメラの前で書かれたひらがなによる空書を認識するための手法を開発した．

評価実験では，50 音（ゐ，ゑを除く）と濁音・半濁音の 71 文字のひらがなを 1 文字ずつ 5 人の被験者に 2 回書いてもらい識別を行った．検証方法には 5 分割交差検証を用いて検証を行った．トレーニングデータとしては 29 人分のデータを集め，補正なしのものと補正を加えたもの，手書き文字データを追加したものの 4 種類を用意した. 71 文字のひらがなを識別させたとき, 補正なしのものは 96%, 補正ありのものは 97%の識別率であった．また，手書き文字データを追加したものは 98%の識別率となり, 効果があることがわかった．

しかしながら，エラー分析を行った結果，識別できる文字と識別できない文字の間で差があることがわかった．酷似したひらがな同士や濁音・半濁音のような画数が多いひらがなは書き方によって誤認することが多いと分かった．今回適用した上方向の線を消すという補正以外にも，軌跡の余分な線を判別することができればより識別精度の向上が見込まれる．

今後の課題としては，まず色特徴量で設定した手の領域判定が挙げられる．今回は，書きはじめ・書き終わりを色手袋の特徴量で設定した．そのため画面の中に入ったと同時に線の描画が開始され，逆に画面の外に出ない限り線の描画は続いていた．このため，文字の始点・終点以外の線が多く含まれることとなった．例えば，特定の手の形や動きをしたときに，書きはじめ・書き終わりを認識できることができれば，画面内の余分な線を少なくすることができると思われる．また，手の自動認識を行うことが可能となれば，色手袋を用いる必要もなくなるため，非接触型の空書認識が可能になると思われる．

また，本手法では，求められた手の特徴量から重心を求め，その重心の軌跡によって文字を再現した．しかしながら，被験者の中には指先で文字を書いているイメージで書いている人が一定数おり，そのため被験者が思っている通りの字になっていないことや乱れることが多かった．これは学習データにも影響することがあるため指先の認識を行うことも今後の検討課題である．

また，本手法では，線の切れ目として，左上に向かうベクトルの線を削除した．これによって人の目では非常に見やすくなり，全体的な識別率も向上した．しかし，文字を識別するのに必要な部分の線も消してしまっており，誤認を起こす元にもなってしまった．そのため，上方向の線を消す以外の手法についても検討する必要がある．手の位置座標から文字を書く際に通らない位置や場所を特定することができれば除外する線がわかりやすくなる．反辞書確率モデルを用いた場合，重心の座標を取得しているため，座標の流れから異常を見つけることができると思われる．

本手法では, CNN による空書文字の識別を行った．画像による識別のため，手書きの文字と酷似していることから手書き文字データの追加を行い，データ量を補うことも行った．結果的には識別率はわずかに上がっていたため，効果はあると思われる．今後は，空書の動画データをより集めることにより，画像による CNN ではなく, 動きを含めた動画による R-CNN による認識についても検討したい．

参 考 文 献

(1) 藤本一文, 長谷川忍 : ”深層学習を用いた単眼カメラによる空書の自動認識”, 2019 年度教育システム情報学会北信越支部学生研究発表会, pp.41-42, (2020).

(2) 浅野敏郎, 宮田明, 本田幸生 : ”空中文字ジェスチャを用いた視覚インタフェース”, 精密工学会誌, pp.333-337, (2011).

(3) 杉本真佐樹, 中井一文, 江崎修央, 清田公保 : ”Wii リモコンを用いた目本語文章の入力方法”, 映像情報メディア学会技術報告, pp.59-62, (2011).

(4) 保呂毅, 稲葉雅幸 : ”複数カメラを用いた手書き文字認識システム”, 第 14 回インタラクティブシステムとソフトウェアに関するワークショップ, (2006).

(5) 熊澤遼, 渡辺亮 : ”Kinect を用いた NUI システムの構築～ジェスチャと指先本数認識を利用した TV の操作～”, 自動制御連合講演会講演論文集, pp.26-28, (2016).

(6) 秦優哉, 小森一誠, 川名晴也, 大枝真一 : ”CNN のアンサンブル学習による文字認識の正誤判定評価”, 情報処理学会第 80 回全国大会講演論文集, pp.707-708, (2018).

(7) 紙徳直生，伊藤大喜，多田晃己，孟林，山崎勝弘：”深層学習を用いた甲骨文字認識”, 情報処理学会第 80 回全国大会講演論文集, pp. 513-514, (2018).
(8) K. He, X. Zhang, S. Ren , J. Sun: “Deep Residual Learning for Image Recognition”, arXiv:1512.03385, (2015).
(9) 文字画像データセット(平仮名 73 文字版)を試験公開しました, https://lab.ndl.go.jp/cms/hiragana73, (2020 年 1 月 16 日確認).