Web カメラを用いた指文字自動認識システム

渡邉 聡* 五十嵐 悠紀*

概要. Web カメラを用いてユーザが示した指文字を自動認識するシステムを提案する．実装には手の形状予測モデルである Handpose と本提案で制作した指文字予測モデルの 2 種類の機械学習モデルを使用した．Handpose を利用することで，背景映像に関わらず手の形状を高精度で取得することが可能である．また，指文字の認識精度を高くするために距離別に取得した画像を用いて機械学習を行うことで認識精度を 90%まで上げることに成功した．

1 はじめに

近年はオンライン化が進み，日々の生活に音声・ビデオ通話がより一層根付いた．こういった発展に従って，会話音声を自動で文字に起こし議事録を作成するシステム等も発展している．このシステムは議事録を作成するだけでなく，耳の聞こえない人も音声通話の内容にリアルタイムで参加する 1 つの方法となっている．また文字を音声にする等の技術もあり，これは目の見えない人が議事録を確認する方法となっている．

しかしながら，何らかの不自由を持っている人から発信するものを変換するシステムはあまり開発されていない．これはタイピングによって文字に起こすことで意思疎通を図ることができるからである．だがタイピングは口頭での会話等に比べて時間がかかり，リアルタイムでの参加にはワンテンポ遅れてしまうこともある．リアルタイムで点字や手話を自動で文字に起こすまたはそれらの学習が容易になれば，相互の意思疎通が容易に図れるものと考えた．

本稿では手話に着目し，手話を自動で文字に直すことを目的とする．また手話を認識するシステムを用いて，学習支援システムの作成を行う．手話は無限に増え続けることから，ひらがなと対応したジェスチャーである，指文字に限定することとした．カラーグローブを用いた指文字の認識[1]は提案されているが提案システムでは素手で認識することを目指した．また 1 つの文字の動作時間は一定ではないことから，本提案では動きの無い 41 種類の指文字を対象として認識を行った．手とカメラの距離における精度変化について調査し，適切なモデルの作成を試みた結果，カメラとの距離の種類「近・中・遠」，これらすべてのデータを親データとして作成するモデルが最も精度が良いことが分かった．

* 明治大学 総合数理学部

2 提案システム

指文字自動認識のシステムを Python3.9 と JavaScript を用いて制作した．また手の形状予測のため MediaPipe Handpose[2]を用いた．MediaPipe Handpose とは，機械学習を利用して画像から 21 点の特徴点を出力するモデルである．以下でシステムについて述べる

2.1 使用条件

本システムでは JavaScript を使用しているため，JavaScript を有効にする必要がある．また学習可能な指文字は，動きが無い 41 種類の指文字に限定し，指文字の認識には右手で提示することとした．

2.2 インターフェース

本システムでは，提示される見本の画像とカメラ映像を見比べて形状を確認しつつ，機械学習によって正誤を判定するものである．

図 1 がシステム画面である．画面左側のひらがなパネルを押すとその指文字のイラスト[3]が表示される．本画像では手のみを写しているが，顔や体が

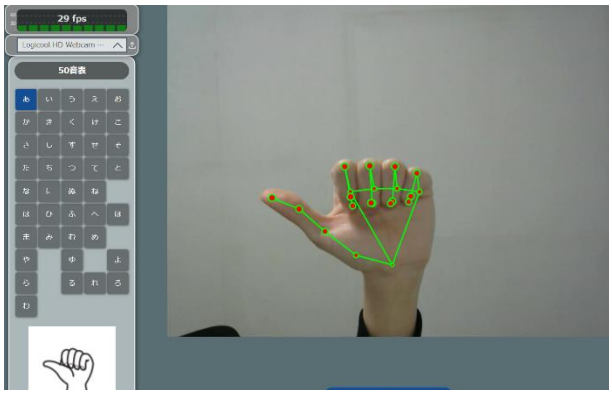

図 1 システムの画面．「あ」の指文字

写っても認識可能である．ユーザが左下に表示されたイラストを真似してカメラ前でハンドジェスチャを行うと，画面下部にある青枠の中の文字が予測結果の文字として表示される．

3　モデル作成手法

指文字認識モデルは Handpose の API から得られる特徴点 21 個の 3 次元座標を使用し，各特徴点の位置関係からモデルの作成を行う．モデル作成手法は Scikit-learn の Support Vector Machine の SVC である．

各指文字に 70 個の教師データを用意し，C=100 γ=0.1 の値で学習を行う．本研究のデータ数の場合，高速に処理可能でありリアルタイムな表示に支障が生じなかったことから，このような値設定とした．

4　モデル比較

本システムで使用するカメラとユーザの手の距離は一定ではない．したがって様々な距離に対応可能な認識モデルを作成する必要がある．本研究における距離の定義は，カメラからの距離が 30cm の「近」，50cm の「中」，80cm の「遠」の 3 種類とし，各距離でデータを収集し，精度の検証を行った．

本節では近・中・遠の三種類の距離で指文字を行い，それぞれの指文字がどの指文字として認識されるか検証する．認識モデルは近・中・遠それぞれの距離で得たデータを教師データとして作成したモデル 3 種類に加え，近・中・遠すべてのデータを教師データとして使用したモデル 1 種類の計 4 種類を使用する．この結果を踏まえて，どのモデルが適切であるか判断する．

4.1 各モデルの精度比較

各指文字を提示した際の推定指文字を図 2 に示す．すべてのモデルにおいて，それぞれの適正距離で高い精度であった．しかしながら，近・中のモデルにおいては適正外の距離における精度は著しく低下した．また，遠距離モデルはすべての距離において精度が比較的高かった．全距離モデルが最も優れており，すべての距離において 90%以上の精度であった．

4.2 処理速度検証

各モデルを使用し，処理速度を計測した．計測する処理時間は，データ操作を行った後から画面上に推定指文字を表示するまでの間である．実験環境はデスクトップ PC であり，ローカルサーバーを用いて POST 通信を行う．また計測した処理時間は，最短処理時間と最長処理時間，そして平均処理時間である．処理回数を 100 回とし，それぞれの処理時間を求める．

計測した結果が表 1 である．最長処理時間が 1ms を超えず，また全距離の処理速度は概ね同様であった．以上からリアルタイム性に問題はないと考える．

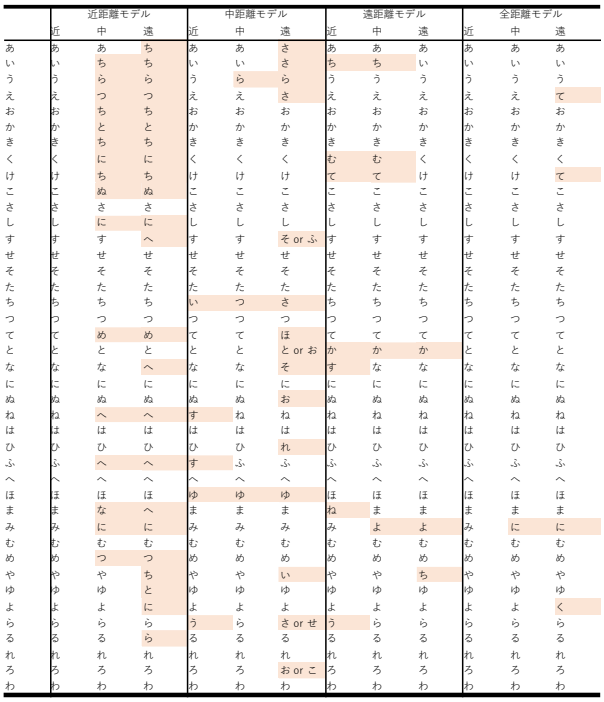

図 2 各モデルの認識結果

表 1 各距離モデル使用時の処理時間(ms)

	全距離	近距離	中距離	遠距離
最小値	0.40	0.40	0.40	0.40
最大値	0.70	0.70	0.80	0.80
平均値	0.54	0.53	0.50	0.58

4.3 結論

各モデルの精度比較を行った結果，全距離のデータを教師データとして作成したモデルが最も精度が高かった．ほかのモデルと比べて教師データ数が増えたが，処理速度は低下せず，他モデルと同様にリアルタイムでの動作に問題は生じなかった．したがって，全距離のデータを教師データとして作成したモデルを使用することが好ましいことが明らかとなった．

参考文献

[1] 渡辺賢，岩井儀雄，八木康史，谷内田正彦．カラーグローブを用いた指文字の認識．電子情報通信学会論文誌 D，Vol.J80-D2，No.10，pp.2713-2722，1997.

[2] “MediaPipeHands” https://google.github.io/mediapipe/solutions/hands.html (2021/11/19 確認)

[3] “ちびむすドリル” https://happylilac.net/sk1805311413.html (2021/11/19 確認)