修士論文

静止指文字と遷移区間を考慮したHMMによる単語認識精度の改善

平成 30 年度修了

三重大学大学院 工学研究科
博士前期課程 情報工学専攻

林 俊明

目次

はじめに

厚生労働省の調査 [1] によると，聴覚障がい者の数は全国で約 34 万人に上る．聴覚障がい者の多くは手話や指文字といったコミュニケーション方法をとっている．指文字とは手話で表現できない固有名詞の表現に使われる視覚言語である．また，昨今においては聴覚障がい者だけではなく，自閉症患者や発達性失語症患者等も手話や指文字を使用するのに対し，手話通訳士 [2] は全国に約 3600 名と少ない．そのため，病院や役所，銀行などで手話や指文字を使う方との円滑なコミュニケーションを支援するシステムが求められている．

手話認識システムはみずほ情報総研が開発し，2013 年に金融国際情報技術展 [3] でデモンストレーションが公開された．指文字認識に関する研究はその多くが静止画像を用いての認識 [4] であったり，動画であっても文字単位の認識 [5] であることが多い．先に述べたように指文字は固有名詞などの表現に用いられるもので，文字単位の認識は実用的ではない．また，[6] はデータグローブを用いて指文字認識を行っている．データグローブとは指の関節部分にセンサーが取り付けられた特殊なグローブで，[6] は認識の際にこのグローブを着用する必要がある．本研究では対象者の負担やコストを削減するためにデータグローブは用いない．[7] や [8] はデータグローブを用いずに単語認識を行っているが，その認識率はそれぞれ 37.4%，54.7% と低い．

指文字の単語認識が難しい理由は，文字から文字へ遷移する区間の途中で別の文字に似た形が出現したり，動きのある指文字と遷移区間の区別が困難であるということが挙げられる．また，英語指文字がアルファベットの 26 種類であるのに比べて，日本語指文字は 82 種類と，カテゴリ数が多いのも認識が難しい理由のひとつである．種類が多いと単に問題として難しいことに加え，学習データの確保が困難となる．なお，手話は動作が大きいため骨格情報などを認識に利用できるが，指文字は動作が小さくほとんど指の動きだけで表現するため骨格情報を利用できない．現状，手指形状推定の精度が低い [9] ため，指文字は手話よりも認識が難しいといえる．

以上のことから，本研究は指文字の単語認識の精度改善を目的とする．認識率の改善目標は現行する音声認識システムと同程度の認識率としたい．株式会社クレスコの調査 [10] によると音声認識サービスの中で最も認識率が高かったのは Google Cloud Speech API で，その認識率は 87.20% であった．

本研究では単語中の静止指文字の認識率を改善するための手法と，削除や挿入誤りを減らすための手法の 2 つを提案する．

本稿では 2.3 節で 1 つ目の手法について，2.5 節で 2 つ目の手法について述べる．

第 1 章
手話と指文字

手話は視覚言語のひとつで，図??のように語彙を表現することができる．指文字は手話の語彙にない固有名詞などの表現に用いられる．指文字は 50 音すべてを表現することが可能で，全 82 種類からなる.「 あ」や「 か」のように手指に動きがない静止指文字と,「 の」や「 ん」のように動きがある指文字がある．また濁音は，例えば「 が」を表現するときは指文字の「 か」を右に移動させて表現する．半濁音は，上に移動，小書き文字は手前に引く動作をする．詳しくは図 1 を参照していただきたい．

図 1　手話と指文字 [11]

第 2 章
従来手法 [7]

2.1 従来手法の流れ

データセットの撮影から認識までの一連の流れについて述べる．

- **撮影**

  撮影には MicroSoft 社の Kinect for Windows(以下 Kinect，図 2) を用いる．Kinect は RGB カメラと深度センサーを内蔵しており，RGB 画像に加えて Depth データを取得できる．照明の影響を統一するために撮影は日没後に行い，ノイズを削減するために被験者は黒色の衣服を着用する．

図 2　Kinect

- **手領域検出**

  入力された動画の各フレーム (図 3) で肌色検索を行う．肌色の条件を式 1 に示す．

$$(R \geq 75) \cap (R + 20 > G) \cap (R + 20 > B) \cap (B \geq 40) \tag{1}$$

  式 1 を満たす肌色領域のうち，重心がもっとも左下にある領域を手領域とする．その後，48 × 48pixel の画像に正規化する (図 4)．

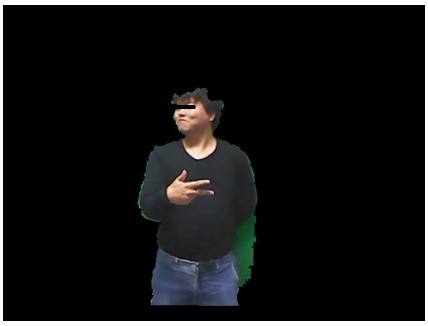

図 3　フレーム画像

図 4　正規化画像

- **特徴量抽出**

  HOG 特徴量とフレーム間での手領域重心の移動座標と移動距離を抽出する．

  – **HOG 特徴量**

  正規化画像から HOG 特徴量を抽出する．輝度勾配は 9 方向に分割し，1 セルを 6 × 6pixel，1 ブロックを 3 × 3 セルとして計算する．入力は 48 × 48pixel の画像であるため，縦方向に 6 ブロック，横方向に 6 ブロック，計 36 ブロックでヒストグラムの正規化を行う．1 ブロックに 9 セル含まれ，輝度勾配が 9 方向であるため，1 ブロックの次元数は 9 セル× 9 方向 = 81 次元となる．よって，得られる HOG 特徴量の次元数は 36 ブロック × 81 次元 = 2916 次元となる．

  – **フレーム間での手領域重心の移動座標と移動距離**

  正規化前の画像からフレーム間での手領域重心の移動座標と移動距離を抽出する．フレーム間での移動座標の単位は

$$\Delta x_{raw}[pixel]$$
$$\Delta y_{raw}[pixel]$$
$$\Delta z_{raw}[mm]$$

  であるため，単位を統一する必要がある．kinect による撮影画像の画角は垂直方向が 43°，水平方向が 57° なので，画像の縦，横の実際の長さ $height_{real}[mm]$，$width_{real}[mm]$ は，以下のように推定できる．

$$height_{real} = 1500 \approx 2 \times 1900 \times \tan \frac{43 \div 2}{180} \pi [mm] \tag{2}$$

$$width_{real} = 2000 \approx 2 \times 1900 \times \tan \frac{57 \div 2}{180} \pi [mm] \tag{3}$$

よって実際の移動座標と移動距離は，

$$\Delta x_{real} = \frac{width_{real}}{480} \times \Delta x_{raw}[mm] \tag{4}$$

$$\Delta y_{real} = \frac{height_{real}}{640} \times \Delta y_{raw}[mm] \tag{5}$$

$$\Delta z_{real} = \Delta z_{raw}[mm] \tag{6}$$

$$\Delta d_{real} = \sqrt{\Delta x_{real}^2 + \Delta y_{real}^2 + \Delta z_{real}^2} \tag{7}$$

また，HOG 特徴量は 81 次元での正規化を行っているため，HOG 特徴量との正規化を行った移動座標と移動距離は以下の様になる．

$$\Delta x_{nor} = \frac{\Delta x_{real}}{9 \times \sqrt{\Delta x_{real}^2 + \Delta y_{real}^2 + \Delta z_{real}^2} + 1} \tag{8}$$

$$\Delta y_{nor} = \frac{\Delta y_{real}}{9 \times \sqrt{\Delta x_{real}^2 + \Delta y_{real}^2 + \Delta z_{real}^2} + 1} \tag{9}$$

$$\Delta z_{nor} = \frac{\Delta z_{real}}{9 \times \sqrt{\Delta x_{real}^2 + \Delta y_{real}^2 + \Delta z_{real}^2} + 1} \tag{10}$$

$$\Delta d_{nor} = \frac{\Delta d_{real}}{81} \tag{11}$$

抽出する特徴量の次元数は，HOG 特徴量の 2916 次元に式 (8)〜(11) の 4 次元を加えた 2920 次元となる．

- **次元削減，シンボル系列生成**

  2920 次元の特徴量を PCA により 200 次元にまで次元削減する，この時，大きい方から 200 個の固有値の累積寄与率は 87.5% 以上 90% 以下となっている．この 200 次元の特徴量の実現値の組をグループ化し，グループごとにシンボルを決め，特徴量の時間変化をシンボル系列として表す．シンボル系列生成には，ベクトル量子化を用いて 250 個の代表値を生成する．

- **学習と認識**

  Hidden Markov Model(以下 HMM) による学習と認識を行う．HMM による学習と認識には HMM Tool Kit(以下 HTK)[12] と呼ばれるソフトウェアを用いる．この HTK に実装されている HInit と HRest で HMM の学習を行い，HVite で認識を行う．

2.1.1 実験

- データセット

被験者 5 名からそれぞれ 2 回ずつ表 1 に示す 64 個の単語を撮影する．表 1 の単語には指文字全 82 種のうち,「 を」を除いた 81 種類の文字が含まれている．

表 1　実験結果

けせんぬま	ねづ	おず	やよい	えにわ	ほくと
きたみ	しべつ	よもぎた	やまだ	ざおう	ごじょうめ
みぶ	かづの	よなぐに	はちのへ	ひゆ	あそ
しゅう	れぢん	あーろん	あじゃ	えいべる	けねでぃ
ばりー	だん	ぶらっど	きゃろる	ちぇすたー	でいう゛
はりそん	ひゅーご	じょなす	れっくす	まーかす	さむ
びーがー	ぼぷつぃん	ふぉくつ	う゛ぁいぜ	ぽぱい	ぎぐ
ざもちん	ぞちこふ	むーみん	ぺがのふ	かいぴお	ぬんめら
わーげん	ゆんほ	さへる	ぢく	どぅーく	ぽんぺい
ぱーぷる	げばら	ちぇ	ぼぁとぅん	ぐっぴー	ずぉーだ
びぜふ	ぞこら	てせ	てしゅ		

- 評価方法

評価式を式 12 に示す．

$$\text{認識率} = \frac{N - D - S - I}{N} \times 100 \tag{12}$$

ただし，N:文字数，D:削除数，S:置換数，I:挿入数である．

正解に対してどのような結果のとき削除，置換，挿入となるのかを表 2 にまとめた．表 2 は a/sil/ru/sil/to を正解としたときの認識結果に対する評価を示している．sil は文字から文字への遷移区間を表す．

表 2　正解に対する評価

認識結果	N	D	S	I	備考
a/sil/ru/sil/to	3	0	0	0	正解
a/sil/to/sil/ru	3	0	2	0	2,3 文字目が置換
ru/sil/a/sil/to	3	0	2	0	1,2 文字目が置換
ru/sil/to/sil/a	3	1	0	1	3 文字目が挿入・ a が削除
to/sil/a/sil/ru	3	1	0	1	1 文字目が挿入・ to が削除
to/sil/ru/sil/a	3	0	2	0	1,3 文字目が置換
a/sil/ru	3	1	0	0	to が削除
a/sil/to	3	1	0	0	ru が削除
ru/sil/a	3	1	1	0	a が削除・ 2 文字目が置換
ru/sil/to	3	1	0	0	a が削除
to/sil/a	3	1	2	0	1,2 文字目が置換・ a もしくは to が削除
to/sil/ru	3	1	1	0	to が削除・ 1 文字目が置換
o/sil/a/sil/ru/sil/to	3	0	0	1	1 文字目が挿入
a/sil/o/sil/ru/sil/to	3	0	0	1	2 文字目が挿入
o/sil/a/sil/to/sil/ru	3	0	2	1	3,4 文字目が置換・ 1 文字目が挿入
a/sil/o/sil/to/sil/ru	3	0	1	1	2 文字目が置換・ 4 文字目が挿入
o/sil/ru/sil/to/sil/a	3	0	1	1	1 文字目が置換・ 4 文字目が挿入
ru/sil/o/sil/to/sil/a	3	0	2	1	1,2 文字目が置換・ 4 文字目が挿入
o/sil/a/sil/ru	3	1	0	1	1 文字目が挿入・ to が削除
a/sil/o/sil/ru	3	0	2	0	2,3 文字目が置換
o/sil/ru/sil/a	3	0	2	0	1,3 文字目が置換
ru/sil/o/sil/a	3	0	3	0	1,2,3 文字目が置換
o/sil/to/sil/a	3	0	3	0	1,2,3 文字目が置換
to/sil/o/sil/a	3	0	3	0	1,2,3 文字目が置換

- **実験結果**

  実験結果を表 3 に示す．表中の H は正解文字数を表す．

表 3　実験結果

	H	D	S	I	N	認識率
1	163	10	259	17	432	33.80
2	188	14	231	16	433	39.72
3	180	9	243	9	432	39.58
4	162	23	247	24	432	31.94
5	183	13	236	16	432	38.66
合計	876	69	1216	82	2161	37.35

第 3 章
提案手法

2.1　遷移区間を個別に学習する

従来手法は音声認識の手法に倣っており，遷移区間を 1 つの HMM で学習している．遷移区間とは図 5 に示したように文字から文字へ移る区間のことである．

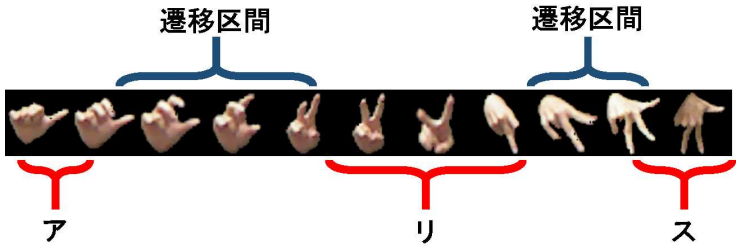

図 5　指文字の遷移区間

図 6　音声の遷移区間

音声認識の場合，遷移区間はどの文字からどの文字への遷移区間であっても無音という同じ特徴を持つことになる (図 6 赤枠)．そのため，音声認識ではどの遷移区間も 1 つの HMM で学習する．従来手法も，この音声認識の手法と同様にどの遷移区間も 1 つの HMM で学習する．

しかし，指文字の単語認識の場合，遷移区間はどの文字からどの文字に遷移するかによって特徴が異なる (図 5)．そのため，従来手法のようにどの遷移区間も 1 つの HMM で学習するのは適切ではない．

そこで，遷移区間の種類ごとに HMM を用意し学習を行った．

2.1.1 実験

- **データセット**

  従来手法 [7] では指文字 81 種類を撮影していた．今回は，限られた時間で十分な学習量を確保するために撮影する文字の種類は 9 種類とした．そして，文字が 9 種類のとき，遷移区間は 9^2 種類存在するが，撮影する遷移区間は 17 種類とした．また，従来手法では被験者 5 名から撮影を行っていたが，個人差を吸収するために今回の被験者数は 10 名とし，表 4 の単語 33 個を各被験者からそれぞれ 1 回ずつ撮影した．撮影条件は従来手法と同じである．

表 4　撮影する単語

ノリ	アリ	アミ	スト	ノコ	ノリオ
アリス	アリア	コリア	リアス	アルミ	アルト
ルミノ	アミノ	コリオ	トリオ	リスト	ノリス
ノコリ	コリス	アスト	アスル	スルミ	オルミ
オルト	スルト	リアミ	トリス	ストリ	ミノコ
リオル	リアル	ミノリ			

- **評価方法**

  従来手法と同じである．

- **実験結果**

  従来手法による実験結果を表 5，遷移区間を個別に学習する手法による実験の結果を表 6 に示す．

表 5　従来手法

	H	D	S	I	N	認識率
1	71	2	21	0	94	75.53
2	81	2	11	10	94	75.53
3	77	9	8	3	94	78.72
4	73	9	12	6	94	71.28
5	72	8	14	7	94	69.15
6	81	3	10	1	94	85.11
7	73	11	10	1	94	76.60
8	78	9	7	0	94	82.98
9	80	1	13	3	94	81.91
10	82	0	12	3	94	84.04
合計	768	54	118	34	940	78.09

表 6　遷移区間を個別に学習した実験の結果

	H	D	S	I	N	認識率
1	87	1	6	2	94	90.43
2	92	0	2	7	94	90.43
3	92	2	0	3	94	94.68
4	92	1	1	6	94	91.49
5	89	3	2	7	94	87.23
6	91	1	2	4	94	92.55
7	88	2	4	1	94	92.55
8	92	0	2	0	94	97.87
9	92	0	2	1	94	96.81
10	92	1	1	2	94	95.74
合計	907	11	22	33	940	92.98

2.1.2 考察

遷移区間を個別に学習する手法によって，認識率は 78.09% から 92.98% に改善された．これは Google Cloud Speech API といった音声認識サービスよりも高い精度である [10]．しかし，今回の実験は 9 種類の文字と 17 種類の遷移区間にしか対応できない．指文字は全部で 82 種類あるため，遷移区間は単純に計算すると 82^2 種類あることになる．実際に撮影を行ってみて，学習に十分な量の遷移区間のデータを全種類撮影するのは現実的ではないと感じた．手形状 CG で人工的に学習データを作るといった研究 [13] もあるが，[13] で作れるデータは静止画のみである．本研究は単語認識を行っており，[13] では手形状がある文字からある文字へ変遷していく動画データは生成できない．

そこで，FantaMorph[14] というソフトでモーフィングという技術を利用して,「ア」から「ミ」への遷移区間を生成してみたがうまくいかなかった (図 7)．

以上のことから，次節以降は遷移区間を個別に学習することなく認識率を改善する手法を提案する．

図 7　モーフィング

2.2 学習量を増やす

表 3 の結果から，文字単位の HMM が正しく学習できていないおそれがあると考えた．正しく学習できていないとすれば，まず考えられるのは学習不足である．そこで，単語認識ではなく文字認識において，学習量の違いによる認識率の変化を確認することにした．

学習量の違いによる認識率の変化を確認するために，文字単位のデータセットを用意した．撮影した文字は「ア，オ，コ，ス，ト，ノ，ミ，リ，ル」の 9 種類の文字で，10 名の被験者からそれぞれ 1 回ずつ撮影した．そして，撮影した動画の各フレームに対して幾何学的変形を施すことによって，抽出される特徴量に変化を与え学習データを増やす．ただし，時間軸に対しては特徴量の変化を与えられていない．増やす量は 11 倍と 53 倍とした．以下に幾何学的変形の種類を示す．

- **11 倍に増量**
  - 元画像 (図 8)
  - 横幅を 0.8，1.2 倍 (図 9，10)
  - 縦幅を 0.8，1.2 倍
  - 左に 5，10 度回転 (図 11)
  - 右に 5，10 度回転 (図 12)
  - せん断因子 0.5，-0.5 で x 軸方向にせん断 (図 13)

- **53 倍に増量**
  - 元画像
  - 横幅を 1.02 倍，1.04 倍，1.06 倍，1.08 倍，1.10 倍
  - 縦幅を 1.02 倍，1.04 倍，1.06 倍，1.08 倍，1.10 倍
  - 左に 0.5 度，1.0 度，1.5 度，2.0 度，2.5 度，3.0 度，3.5 度，4.0 度，4.5 度，5.0 度回転
  - 右に 0.5 度，1.0 度，1.5 度，2.0 度，2.5 度，3.0 度，3.5 度，4.0 度，4.5 度，5.0 度回転
  - せん断因子 0.05，0.10，0.15，0.20，0.25，0.30，-0.05，-0.10，-0.15，-0.20，-0.25，-0.30 で x 軸方向にせん断
  - 1.02 倍，1.04 倍，1.06 倍，1.08 倍，1.10 倍に拡大
  - 0.98 倍，0.96 倍，0.94 倍，0.92 倍，0.90 倍に縮小

図 8　元画像

図 9　横幅を 0.8 倍

図 10　横幅を 1.2 倍

図 11　左に 5 度回転

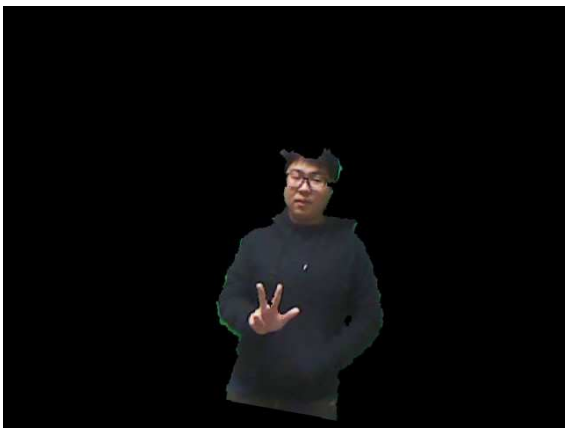

図 12　右に 10 度回転

図 13　剪断係数 0.5 で x 軸方向に剪断

2.2.1 実験

- **評価方法**

  評価式を式 13 に示す．

$$\text{認識率} = \frac{H}{N} \times 100 \tag{13}$$

  ただし，H:正解数，N:文字数である．

- **実験結果**

  学習量を 11 倍に増やした実験の結果を表 7，53 倍に増やした実験の結果を 8 に示す．表中の文字は認識結果である．

表 7　学習量を 11 倍した実験の結果

	1	2	3	4	5	6	7	8	9	10	正解数
ル	ノ	ノ	リ	ル	リ	コ	ル	リ	ノ	オ	2
コ	ノ	コ	ノ	ノ	コ	コ	ノ	コ	ノ	コ	5
オ	オ	ノ	リ	リ	ト	オ	オ	リ	ノ	ノ	3
ア	コ	ノ	リ	コ	ト	リ	ノ	ア	ノ	ア	2
ト	リ	コ	ノ	ト	リ	ノ	ノ	リ	ノ	ノ	1
ス	リ	ト	リ	ス	ト	ノ	ノ	ノ	ス	ス	3
ミ	ノ	ミ	ミ	ミ	ミ	ミ	ミ	ミ	リ	ミ	8
ノ	ノ	ノ	ノ	ノ	ノ	ノ	ノ	ノ	ノ	ノ	10
リ	ノ	リ	リ	リ	リ	ノ	リ	リ	ノ	リ	7
正解数	2	4	3	6	4	4	5	5	2	6	41

表 8　学習量を 53 倍した実験の結果

	1	2	3	4	5	6	7	8	9	10	正解数
ル	コ	ノ	ル	ル	ト	ト	コ	ノ	オ	ト	2
コ	ル	ノ	ト	ア	ト	コ	ス	ノ	オ	コ	2
オ	ト	ア	ト	オ	ル	オ	オ	ル	オ	オ	5
ア	ル	ア	ア	ア	ア	ト	オ	ト	ア	コ	5
ト	ト	ト	ミ	ト	オ	コ	コ	ノ	ト	ト	5
ス	オ	ス	ト	ス	ス	コ	ス	ル	ス	ス	6
ミ	ル	ミ	ト	ア	ミ	ミ	コ	ミ	ミ	ミ	6
ノ	ノ	ノ	リ	ノ	ノ	ノ	コ	ノ	ノ	コ	7
リ	ノ	リ	リ	リ	ト	ト	ノ	リ	ノ	リ	5
正解数	2	6	3	7	4	4	2	3	6	6	43

2.2.2　考察

学習量を 11 倍した実験の認識率は 45.56%，53 倍は 47.78% となった．本節の実験で学習に使われた動画の数は，11 倍の場合は 1 文字あたり 99 個，53 倍の場合は 477 個である．学習量が大きく変わったにも関わらず，認識率はわずかな変化であった．

実験に用いた文字のうち「ア，オ，コ，ス，ト，ル」は静止指文字,「ノ，リ」は動きのある指文字である (図 1)．静止指文字の認識率は表 7，8 でそれぞれ，34.29%，44.29% であるのに対し，動きのある指文字はそれぞれ 85.0%，60.0% であった．この結果から，静止指文字は動きのある指文字よりも認識率が低いことが分かった．次節では静止指文字の認識率を改善する手法を提案する．

また，以降の節では 1 文字あたり 100 個程度の学習量で実験を行うこととする．

2.3 静止指文字に時間的な特徴の変化を与える

前節では，動きのある指文字より静止指文字の認識率が低いことが分かった．2.1.1 の実験結果についても静止指文字と動きのある指文字の認識率の違いを確認した．その結果を表 9 に示す．

表 9　2.1.1 の静止指文字と動きのある指文字の認識率

	静止指文字		動きのある指文字	
	正解数	認識率	正解数	認識率
1	57	25.91	106	50.00
2	72	32.73	116	54.72
3	79	35.91	101	47.64
4	71	32.27	91	42.94
5	69	31.36	114	53.77
合計	348	31.64	528	49.81

表 9 からも動きのある指文字より静止指文字の認識率が低いことが分かる．このような結果には以下のような原因が考えられる．

本研究では従来手法と同様に識別器に HMM を使っている．HMM は特徴量の時間的変化を学習する確率モデルであり，その性能は音声認識で確かめられている [15]．しかし，静止指文字の場合，特徴量が時間的にほとんど変化せず，ブレやノイズによる微妙な変化があるのみである．HMM は先に述べたように特徴量の時間的変化を学習する確率モデルである．そのため，HMM で静止指文字を学習しようとすると，わずかなブレやノイズに反応し，その変化がその文字特有の時間的変化であると学習してしまう．その結果，静止指文字の誤認識が起きていると考えられる．

これを解決するために，本節では静止指文字に対して時間的な特徴の変化を与えることにした．

具体的には入力動画に対し幾何学的変形を施すことによって，静止指文字に時間的な特徴の変化を与える．しかし，その方法には 3 つの制限がある．まず，動きのある指文字の認識に悪影響を与えてはいけない．例えば，静止指文字に時間的な特徴の変化を与えるために，差分と移動平均を用いて画素値が下がっていくような加工をしたとする (図 14，15)．

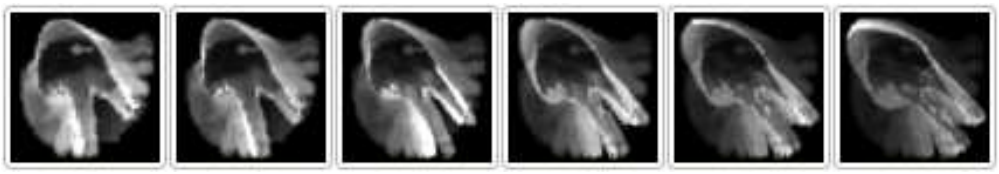

図 14　画素値が下がっていくように加工した動きのある指文字

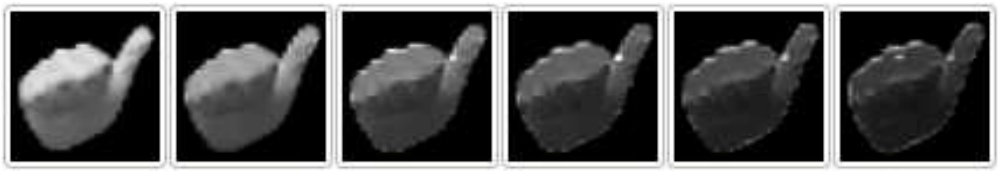

図 15　画素値が下がっていくように加工した静止指文字

図 14 を見てわかるように，動きのある指文字にこの加工をすると移動の軌跡が現れてしまい，手形状が複雑になることで認識に悪影響を及ぼす恐れがある．また，単語認識に応用することを考え，繰り返し適用できる加工でないといけない．単語認識になると動画のフレーム数が増えるため，この加工方法だといずれ黒画像になってしまうため認識ができなくなる (図 15)．さらに，加工によってその文字が別の文字に類似してはいけない．

以上のことを踏まえ，以下のように加工することにした．

加工方法 1　1 フレームおきに黒画像に差し替える
加工方法 2　フレーム内で手領域重心を水平に-3〜3 まで 1 ピクセルずつ往復移動させる
加工方法 3　1 フレームおきに手領域を 0.8 倍する

加工前の動画の各フレーム例を図 16，加工方法 1〜3 で加工した動画を図 17〜19 に示す．

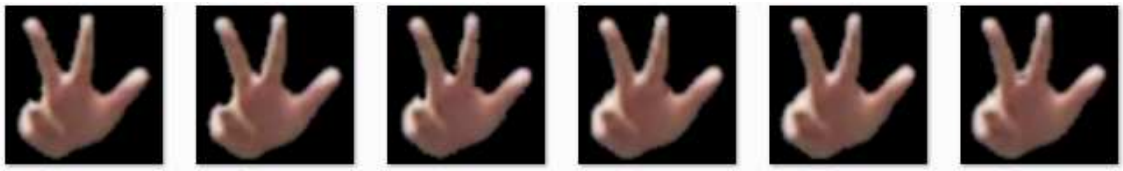

図 16　元画像

図 17　加工方法 1

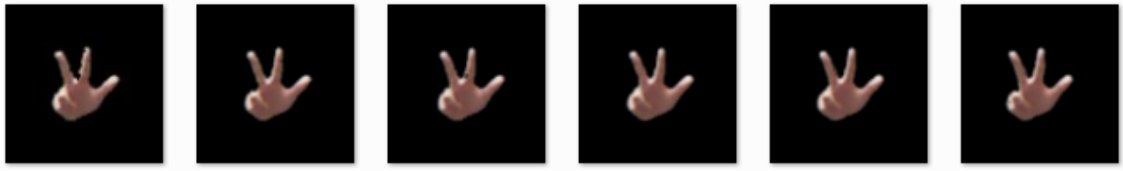

図 18　加工方法 2

図 19　加工方法 3

なお，加工方法 3 は，指文字の小書き文字を表現する際の引く動作と同じ見え方をすると思われるかもしれない．これについては，引く動作の場合は手領域がフレームを追うごとにだんだん小さくなっていくのに対して，加工方法 3 は 1 フレームおきの縮小であるため見え方は異なる．また，以上の加工は深度データまでは加工しない．小書き文字の深度データはフレームを追うごとに変化するのに対して，加工方法 3 で加工した静止指文字の深度データは変化しない．以上のことから，加工方法 3 によって加工した文字とその小書き文字は類似しない．

2.3.1 実験

- **データセット**

  前節で学習量を 11 倍したデータセットを用いる．

- **評価方法**

  前節と同じである．

- **実験結果**

  動画の加工をしなかった場合の実験結果を表 10，加工方法 1〜3 で加工した場合の実験結果を表 11〜13 に示す．これまで行ってきた実験と同様に 10 交差検証を行ったが，表 10〜13 はそのマイクロ平均値をとってある．

表 10　動画の加工なし

静止指文字	動きのある指文字	全体
34.2%	85.0%	45.6%

表 11　1 フレームおきに黒画像に差し替え

静止指文字	動きのある指文字	全体
67.1%	60.0%	65.6%

表 12　-3〜3 まで往復

静止指文字	動きのある指文字	全体
80.0%	90.0%	82.2%

表 13　1 フレームおきに 0.8 倍

静止指文字	動きのある指文字	全体
95.7%	90.0%	94.4%

2.3.2 考察

表 13 の認識率が最も高い結果となった．表 10 と比較して，表 12，13 はいずれも動きのある指文字の認識率に悪影響を与えなかった．表 12 よりも表 13 の方が認識率が高い理由は，表 13 はフレーム間の差分が手領域の輪郭を表すようになるため，より文字固有の特徴を表しやすかったからであると考えられる．表 11〜13 の中で表 11 の認識率が最も低いのは，黒画像が文字固有の特徴をもたないからであると考えられる．

2.4 前節の手法で単語認識を行う

前節で提案した，静止指文字に時間的な特徴の変化を与える手法で単語認識を行った．

2.4.1 実験

- データセット

従来手法の実験では全種類の指文字の動画を撮影していた．今回の実験では限られた時間で十分な学習量を確保するために撮影する文字の種類を減らし，1 文字辺りの動画数を増やすことにした．指文字が全部で 82 種類ある中で，どの文字を撮影対象に選ぶかは以下のように決めた．

まず，全指文字をその動きごとに分けると表 14 のようになる．

表 14　全指文字の内訳

文字の種類	動作	数
静止指文字	静止	41
形状が変化する指文字	文字により異なる	6
濁音	右	21
半濁音	上	5
小書き文字	手前に引く	9
合計		82

また，認識が難しい文字についても考える．指文字全 82 種類のうち形が似ている文字同士の組は以下の 14 組である．

（ア，サ），（イ，キ，チ，ツ，メ），（ウ，セ，ト，ヒ），（カ，ラ），
（ク，コ），（ケ，テ，ホ），（シ，ム），（ス，フ，ネ），（ナ，マ），
（ニ，ミ，ヨ），（ユ，ワ），（ル，レ），（ヌ，ロ），（ノ，リ，長音）

そして，動きは異なるが形がまったく同じ文字同士の組は以下の 28 組である．

（ア，ァ），（イ，ィ），（ウ，ゥ，ヴ），（エ，ェ），（オ，ォ，ヲ），（カ，ガ），
（キ，ギ），（ク，グ），（ケ，ゲ），（コ，ゴ），（サ，ザ），（シ，ジ），（ス，ズ），
（セ，ゼ），（ソ，ゾ），（タ，ダ），（チ，ヂ），（ツ，ヅ，ッ），（テ，デ），
（ト，ド），（ハ，バ，パ），（ヒ，ビ，ピ），（フ，ブ，プ），（ヘ，ベ，ペ），
（ホ，ボ，ポ），（ヤャ），（ユ，ュ），（ヨ，ョ）

表 14 と同じ割合になる様に撮影する文字を選んだ．選んだ文字を表 15 に示す．

表 15　撮影する文字

文字の種類	動作	文字	数
静止指文字	静止	ア，オ，コ，サ，ス， チ，ト，ヒ，ヘ，ミ， ヤ，ヨ，ル	13
形状が変化する指文字	文字により異なる	ノ，リ	2
濁音	右	ゴ，ザ，ジ，ズ，ヂ，ド，ベ	7
半濁音	上	ピ，ペ	2
小書き文字	手前に引く	ャ，ョ，ッ	3
合計			27

そして，選ばれたこれらの文字は前述した認識が難しい文字についても，含まれる割合が同じになるようにした．撮影するデータセットに含まれる文字で，形が似ている文字の組は以下の 5 組である．

（チ，ツ），（ト，ヒ），（ノ，リ），（ミ，ヨ），（ア，サ）

そして，動きは異なるが形がまったく同じ文字の組は以下の 9 組である．

（コ，ゴ），（サ，ザ），（ス，ズ），（チ，ヂ），（ト，ド），（ヒ，ピ），
（ヘ，ベ，ペ），（ヤ，ャ），（ヨ，ョ）

表 16 に撮影する単語 99 個を示す．

表 16　撮影する単語

ノリ	アリ	アミ	スト	ノコ	サジ
チズ	チャ	ジョ	ヨゴ	ヤベ	ヒジ
ゴサ	ヤジ	ドジ	リアル	ミノリ	ノリオ
アリス	アリア	コリア	リアス	アルミ	アルト
ルミノ	アミノ	コリオ	トリオ	リスト	ノリス
ノコリ	コリス	アスト	アスル	スルミ	オルミ
オルト	スルト	リアミ	トリス	ストリ	ミノコ
リオル	ドッチ	ピッチ	ドッジ	ドッヂ	ベッド
ヘッド	ゴッド	ジッチ	ヘッジ	ザッピ	サチヨ
チヅヨ	ジャズ	ヒサヨ	ヤチヨ	ベチョ	ペチャ
ベチャ	ベサヘ	ヤドヤ	ヘヂャ	ドゴヘ	ヨピャ
ヂョザ	ヂョサ	ペピョ	ペヘヨ	ズヨヒ	ベヤザ
ヒャヘ	ドゴヒ	ヒョヒ	ゴベゴ	ヒョズ	ザヤヒ
ヤヂャ	ヂピャ	ピョズ	ジョピ	ズヘズ	ズベザ
ペサザ	ズサゴ	ベザヨ	ヤペザ	ペヂペ	ベザサ
ピャペ	ペゴヘ	ヘサヂ	ヨヂャ	ヒピョ	ペヂヤ
サズゴ	ペチョ	ビャッ			

これらの単語を被験者 20 名それぞれから撮影し，2 名を 1 組とした 10 交差検証を行う．撮影条件は従来手法と同じである．

- **評価方法**

  従来手法と同じである．

- **実験結果**

  従来手法による実験結果を表 17，静止指文字に時間的な特徴の変化を与える手法による実験結果を表 18 に示す．またそれぞれの手法で静止指文字と動きのある指文字の認識率を表 19，20 に示す．

表 17　従来手法

	H	D	S	I	N	認識率
1	160	9	111	8	280	54.29
2	155	18	107	14	280	50.36
3	184	9	87	5	280	63.93
4	156	34	90	4	280	54.29
5	156	9	115	7	280	53.21
6	158	22	100	5	280	54.64
7	133	14	133	10	280	43.93
8	160	16	104	3	280	56.07
9	162	12	106	5	280	56.07
10	176	18	86	4	280	61.43
合計	1600	161	1039	65	2800	54.82

表 18　静止指文字に時間的な特徴の変化を与える手法

	H	D	S	I	N	認識率
1	194	5	81	12	280	65.00
2	177	5	98	8	280	60.36
3	192	9	79	12	280	64.29
4	174	31	75	23	280	53.93
5	200	4	76	10	280	67.86
6	196	9	75	5	280	68.21
7	152	24	104	16	280	48.57
8	173	15	92	13	280	57.14
9	165	8	107	9	280	55.71
10	193	16	71	10	280	65.36
合計	1816	126	858	118	2800	60.64

表 19　従来手法

	静止指文字		動きのある指文字	
	正解数	認識率	正解数	認識率
1	55	43.65	105	68.18
2	68	53.97	87	56.49
3	60	47.62	124	80.52
4	55	43.65	101	65.58
5	76	60.32	80	51.95
6	65	51.59	93	60.39
7	58	46.03	75	48.70
8	59	46.83	101	65.58
9	57	45.24	105	68.18
10	75	59.52	101	65.58
合計	628	49.84	972	63.12

表 20　静止指文字に時間的な特徴の変化を与える手法

	静止指文字		動きのある指文字	
	正解数	認識率	正解数	認識率
1	63	50.00	131	85.06
2	61	48.41	116	75.32
3	80	63.49	112	72.73
4	55	43.65	119	77.27
5	93	73.81	107	69.48
6	80	63.49	116	75.32
7	73	57.94	79	51.30
8	70	55.56	103	66.88
9	68	53.97	97	62.99
10	82	65.08	111	72.08
合計	725	57.54	1091	70.84

2.4.2 考察

静止指文字に時間的特徴を与える手法によって認識率が 54.82% から 60.64% に改善された．表 19，20 を見ると，静止指文字の認識率は 49.84% から 57.54% に改善されており，この手法の有効性が確かめられる．動きのある指文字も見掛け上認識率が改善されているが，t 検定の結果によると平均値に差がない可能性が高い．静止指文字については有意水準 5% で平均値に差があるとの結果が得られた．しかし，置換誤りの数は 1039 から 858 に減ったが，削除と挿入誤りの数は減らなかった．

2.5 特徴量を追加する

表 17，18 を比較してわかるように，提案手法によって置換数は 1039 個から 858 個に減少した．しかし，削除と挿入の数の合計は 226 個から 244 個に増加している．削除と挿入は，文字を遷移区間であると誤認識したり，遷移区間の途中で文字を認識してしまうことによって起こる．つまり，従来手法や静止指文字に時間的な特徴の変化を与える手法では文字と遷移区間を正しく区別できていないということになる．これは，抽出している特徴量に原因があると考えた．

まず，指文字と遷移区間を区別できるような特徴を挙げたものを表 21 に示す．区別するために挙げた特徴量は形状変化と移動方向，移動距離である．形状変化は，その文字または遷移区間を表現している間の手領域の形状変化の有無を表す．移動方向は，動作主から見た場合の手領域が移動する方向を表す．移動距離は動作主を正面から見た場合の手領域の移動距離を表す．形状特徴は，遷移区間全種類のデータを撮影し学習することが困難なため省いてある．

従来手法で抽出する特徴量は HOG 特徴量と移動座標，移動距離であった．つまり，表 21 のうち移動距離しか抽出できていない．その場合，例えば認識の際に移動しない区間を見つけたとき，それが静止指文字なのか遷移区間なのかを区別できていないことになる．

表 21　指文字 (上) と遷移区間 (下) の特徴

	形状変化	移動方向	移動距離
静止指文	無	移動なし	0
小書き文字，を	無	手前	0
濁音	無	右	1 以上
半濁音	無	上	1 以上
長音，も	有	下	1 以上
の，り	有	左下	1 以上
ん	有	下→右上	1 以上
静止指文字，半濁音，小書き文字，を，ん→任意の文字	有	移動なし	0
濁音→任意の文字	有	左	1 以上
長音，も→任意の文字	有	上	1 以上
の，り→任意の文字	有	右上	1 以上

そこで，本節では従来手法で抽出している移動距離に加え，形状変化率と移動方向も抽出することにした．形状変化率と移動方向はそれぞれ式 (14)，(15) で求める．式 (14) はゼロ平均正規化相互相関関数とよばれており，テンプレートマッチングにおける類似度の計算に用いられる関数である [16].

$$\frac{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}((I(i,j)-\bar{I})(T(i,j)-\bar{T}))}{\sqrt{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}(I(i,j)-\bar{I})^2\times\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}(T(i,j)-\bar{T})^2}} \tag{14}$$

N，M　：それぞれ画像の縦と横の長さ

$I(i,j)$，$T(i,j)$　：それぞれ現フレームと前フレームの座標 (i,j) の値

$\bar{I}$，$\bar{T}$　：それぞれ現フレームと前フレームの値の平均

$$\tan^{-1}\frac{d_y}{d_x}\times\frac{180}{\pi} \tag{15}$$

d_x，d_y　：前フレームから現フレームへの手領域重心の移動量

2.5.1 実験

- **データセット**

  前節と同じである．

- **評価方法**

  従来手法と同じである．

- **実験結果**

  静止指文字に時間的な特徴の変化を与える手法による実験の結果を表 22(再掲)，特徴量を追加する手法で行った実験の結果を表 23 に示す．

表 22　静止指文字に時間的な特徴の変化を与える手法 (再掲)

	H	D	S	I	N	認識率
1	194	5	81	12	280	65.00
2	177	5	98	8	280	60.36
3	192	9	79	12	280	64.29
4	174	31	75	23	280	53.93
5	200	4	76	10	280	67.86
6	196	9	75	5	280	68.21
7	152	24	104	16	280	48.57
8	173	15	92	13	280	57.14
9	165	8	107	9	280	55.71
10	193	16	71	10	280	65.36
合計	1816	126	858	118	2800	60.64

表 23　特徴量を追加する手法

	H	D	S	I	N	認識率
1	187	8	85	9	280	63.57
2	182	14	84	11	280	61.07
3	178	9	93	8	280	60.71
4	181	18	81	16	280	58.93
5	198	15	67	15	280	65.36
6	200	6	74	4	280	70.00
7	157	29	94	18	280	49.64
8	164	16	100	13	280	53.93
9	166	11	103	14	280	54.29
10	203	8	69	15	280	67.14
合計	1613	134	850	123	2800	60.46

2.5.2　考察

削除と挿入誤りの数を減らすことができなかった．原因としては，新たに追加した特徴量が主成分分析の際に消失してしまったのではないかと考えられる．そのため，移動距離・形状変化率・移動方向は主成分分析をしたあとの特徴量に付与するべきだった．

おわりに

静止指文字に時間的な特徴の変化を与える手法によって，指文字の認識率が 54.82% から 60.64% に改善された．このことから，HMM による動画像の認識において，動画中の静止物体に時間的な特徴を与えることで認識率が改善されることがわかった．しかし，認識率は依然として低いままである．今後は，入力する動画に対して幾何学的変形を施し時間的な特徴を与えるのではなく，抽出した特徴量に直接変更を加えて静止指文字に時間的な特徴を与えることでさらなる認識率の改善を期待できると考えている．

また，本研究では認識に対応する文字が 27 文字に限定されている．これは指文字全 82 種類の 3 分の 1 にしか満たない．今後は引き続きデータセットを撮影し，全種類の指文字に対応できるようにするべきである．

謝辞

日ごろから多くの御指導を頂きました太田義勝教授，鈴木秀智准教授に深く感謝いたします．そして，日頃何かとお世話になりました落合美子事務員に感謝いたします．また，本論文作成にあたって特にお世話になりました鈴木秀智准教授に深く感謝いたします．最後に，日頃から熱心に討論して頂いた研究室の諸氏に感謝いたします．

参考文献

[1] 厚生労働省社会・援護局障害保健福祉部企画課，“生活のしづらさなどに関する調査”，平成30年

[2] 手話通訳士名簿，http://www.jyoubun-center.or.jp/slit/list/

[3] FIT2018，https://fit-tokyo.nikkin.co.jp/

[4] 慶島淳一，“距離画像を用いた決定木による指文字認識”，システム制御情報学会論文誌，Vol.19，No.4，pp.166-168，2006

[5] 三宅太一，“距離画像を用いた動きのある指文字認識に関する研究”，筑波技術大学大学院，平成24年度

[6] 井出英人，“データグローブを用いた手振り認識システム”，電学論C，112巻5号，平成4年

[7] 大野雄士朗，“HMMによる日本語指文字の動きを考慮した単語認識”，三重大学大学院，平成27年度

[8] 高橋遼平，“Kinectを用いたHMMによる連続指文字認識の検討”，情報処理学会研究報告，Vol.2016-AAC-1，No.9，2016

[9] 山崎祐，星野聖，“指領域情報を用いた手指形状推定”，電子情報通信学会東京支部，平成25年

[10] クラウド型音声認識を評価。精度が良いのはどれ？ https://www.cresco.co.jp/blog/entry/4360/

[11] 千歳市社会福祉協議会，https://www.chitose-shakyo.or.jp/archives/10906.html

[12] htk，http://htk.eng.cam.ac.uk/

[13] 田路賢太郎，“指文字識別における手形状CGを用いた学習”，筑波大学，2011

[14] FantaMorph，http://www.fantamorph.com/jp/

[15] “隠れマルコフモデルによる音声認識と音声合成”，徳田恵一，IPSJ Magazine Vol.45 No.10，2004，p.1011

[16] 画像処理ソリューション，http://imagingsolution.blog.fc2.com/blog-entry-186.html

付録

1　作成したプログラムおよび実験データについて

- 実験用ディレクトリ

  /home/t-hayasi/research/t-hayasi_research/oni/test/test/test/OniPlayer

  　撮影プログラム

  /mnt/fs4/t-hayasi/M2/0808tangoagain

  　従来手法の実験のためのディレクトリ (従来手法で撮影されたデータセットを使用)

  /mnt/fs4/t-hayasi/M2/1203_2018test

  　従来手法の実験のためのディレクトリ (本研究で撮影したデータセットを使用)

  /mnt/fs4/t-hayasi/M2/1314tangokakusyuku

  　提案手法 1 の実験のためのディレクトリ

- 各実験用ディレクトリ内プログラム

  main/body/main_do_0.cpp

  　特徴量抽出用プログラム

  pca/t_main_1_pca.cpp

  　主成分分析用プログラム

  main/body/main_do_1.cpp

  　特徴量ファイル作成用プログラム

  HQuant.cpp

  　シンボル系列生成用プログラム

  HInit.cpp

  　学習，認識用プログラム

- データセットディレクトリ

  /home/A_OLDUSR/2015/ohno/opencv/research/ohno_research/CapturedFrames/Word

  　従来手法で撮影されたデータセット

  /mnt/fs4/t-hayasi/2018dataset

  　本研究で撮影したデータセット