Vol.2016-AAC-1 No.9
2016/7/30

Kinect を用いた HMM による連続指文字認識の検討

高橋遼平[†1]　堀内靖雄[†1]　川本一彦[†1]　下元正義[†2]
眞崎浩一[†2]　黒岩眞吾[†1]　鈴木広一[†2]

概要：本論文では Kinect と隠れマルコフモデル(HMM)を用いた連続指文字認識の手法について検討する．先行研究では局所特徴を用いて指文字手型単体の認識を行っていた．しかし指文字とは本来連続で行われるものであるため，連続認識手法を検討する必要がある．そこで本研究では，Kinect から得られる Depth 手型画像に対して局所特徴である Histograms of Oriented Gradients(HOG)を計算し，その時系列データから HMM を使用して連続指文字認識を行う手法を提案する．評価実験の結果，特定話者の認識精度がデータクローズで 95.4%，データオープンで 93.0%，不特定話者の認識精度が手話者クローズで 87.4%，手話者オープンで 54.7%となった．

キーワード：指文字認識，Kinect，隠れマルコフモデル，HOG

1. はじめに

近年，ろう者の社会進出に伴い，様々な場面でコミュニケーションをとる機会が増えている．しかしながら，ろう者と手話を知らない聴者がコミュニケーションをとるとき，いくつかの問題が存在する．一つは筆談を用いる場合である．手話は日本語と文法が異なるため，手話を母語とするろう者にとっては，日本語を書くことは負担となり，コミュニケーション速度が遅くなってしまう．一方，手話通訳者を介する場合，手話通訳者の数が限られている上に，守秘義務を伴う通訳者であっても，プライバシーに関わることであれば話しづらいことがある．そこでこれらの問題を解決し，円滑なコミュニケーション支援を目的とした手話翻訳システムの開発が望まれている．手話翻訳システムとは入力された手話に対し，それに対応した適切な日本語へ翻訳し，出力するシステムとなっている．一方，手話では外来語等，手話表現が存在しない単語を表現したい場合「指文字」を使用する．指文字とは，外来語や地名，人名などの単語を表現するときに用いられる手話で，かな文字それぞれを特定の手型で表現する．そこで本研究では手話で使用される「指文字」に着目し，システムの入力部に当たる認識手法の検討を行う．

2. 先行研究と目的

指文字認識の従来手法として接触型と非接触型がある．データグローブなどを用いる接触型は，動きの検出精度が高く，処理しやすいことから先行研究でも多く用いられてきた[1][2]．しかし，装置を直接身体に装着しなければならないため煩わしく，ユーザへの負担が大きいという問題がある．また，装置が高価であるため，一般のユーザは入手が困難である．

一方，ビデオカメラから得られる動画像を用いる非接触型ではデバイスを身体に装着しないため，煩わしさや手の動作範囲の制限はない．しかし，一般的なビデオカメラを用いた動画像による 2 次元情報では手の前後方向の動きの識別が難しいという問題や，肌色情報を用いるため，背景の色の制限が必要であるという問題，手が顔や反対の手と重なってしまう場合，片手だけを抽出して検出することが困難であるというような問題があった[3]．そこで，ステレオカメラや距離計測カメラと組み合わせることで，手型を抽出し，認識に用いる手法が検討されている[4]．

指文字手型の認識の際，用いる特徴量として，画像局所特徴を用いる手法[5][6]や，爪や手の輪郭線を用いた手法[7][8]も提案されている．また，上記 2 つを組み合わせた手法も提案されている[9]．しかしながら，これらは指文字手型単体で行われるものであり，連続での認識，または動きを伴う指文字認識に対応していない．一方で，動きを伴う指文字の認識手法も提案されている[10][11]．[10]は胸に付けたカメラで自身が行った指文字を撮影し，その軌跡から動きのある指文字（「の」「も」「り」「ん」）を認識する．しかし，ユーザ自身がカメラを着用しなければならないので負担が大きく，また腕の動きを伴う指文字（濁音，半濁音）には対応していない．[11]は腕の動きを伴う指文字（濁音，半濁音）の認識は行えるが，手型自体が動く指文字（「の」「も」「り」「ん」）には対応しておらず，また単語としての連続認識にも対応していない．

本研究では画像局所特徴に着眼し，Microsoft 社の Kinect v2 を用いた非接触型の手法を用いる．先行研究では指文字手型単体の認識であった．しかし一連の手話文を認識する際，その文中で表現された指文字単語を認識可能とすることを目指しているため，連続手話認識に組み込める手法を検討する必要がある．また，動きのある指文字（「の」「も」「り」「ん」濁音，半濁音，拗音，促音，長音）全てを認識対象とする必要がある．そこで本研究では手型特徴量の時系列データから隠れマルコフモデル（以下 HMM）を用いて連続指文字認識を行うことを目的とする．

†1 千葉大学
Chiba University
†2 みずほ情報総研（株）
Mizuho Information ＆ Research Institute, Inc.

3. 提案手法

本研究では Kinect で撮影した指文字動画から距離情報を用いて，各かな文字に対応する手型画像を抽出する．手型画像はブレが少なく，照明変化にロバストな Depth 画像を用いる．次に特徴量として，動きのある指文字（濁音，半濁音など）に対応するために手の位置差分ベクトル⊿を，手型画像の識別のために Histograms of Oriented Gradients（以下 HOG）[12]を使用する．HOG 特徴量については主成分分析により次元圧縮したものを用いる．また，連続で認識を行うため，HMM による学習・認識を行う．

3.1 手型画像の取得

本研究において，指文字の識別を行うためには鮮明な手型のみの画像が必要である．しかしながら RGB 動画像ではブレが大きく鮮明な手型を取得することは困難であるため，ブレが少ない Depth 動画像を用いる．また Kinect による手首座標の取得が失敗することがあるが，本研究では手首座標はうまく取得できたと仮定し，その後の指文字の連続認識を研究の中心とする．そこで今回の収録では，手首の切り取りの処理を簡略化するために被験者の手首に図 1 のような黒いゴムベルトを装着した．

図 1　黒いゴムベルト

手型のみの画像を取得するために，まず，指文字は顔付近で行われるものと仮定し，Kinect から得られる頭座標を用いて顔付近の画像を切り出す．そして背景を切り取るために Kinect から得られる距離情報を利用し，Kinect から 1.1m 後ろを切り取る．（図 2）

次に，Depth 画像に対応した赤外線画像に対して 2 値化処理を行い，手と腕の切り離しをする．そしてラベリング処理を施し，Kinect から得られる手の中心位置がある画像のみを残すことで，残った顔領域や腕領域の除去をする．（図 3）

2 値画像の残された白画素に赤外線画像に対応した Depth 画像をはめ込むことで手型のみの画像を得る．（図 4）

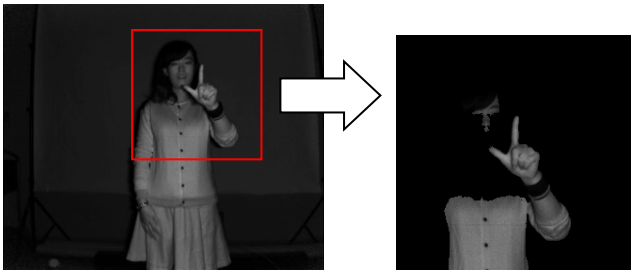


図 2　手領域抽出・背景削除

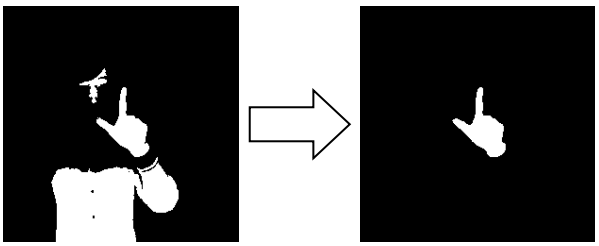


図 3　ラベリング処理

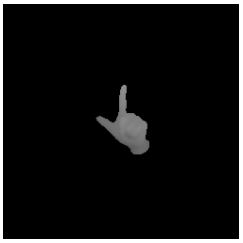


図 4　Depth 手型画像の取得

3.2 特徴抽出

(1) 位置差分ベクトル⊿

得られた手型画像に対し，手の位置座標を抽出する．Kinect から得られる手の位置座標が安定しないときを考慮し，ここでは手型画像の重心を x，y 座標，Kinect からの距離が最も近いピクセルを z 座標とした．また，単語ごと，あるいは手話者ごとでの座標の差をなくすため，各軸のフレーム間差分⊿を特徴量として用いる．

(2) HOG 特徴量

得られた手型画像に対し，HOG を抽出する．しかしながら，HOG は大きさや位置の変化に弱い．そこで手話者間の手の大きさの違いや呈示された指文字の位置の違いを小さくするため，本研究では手型の外接矩形を用いる．求められた外接矩形の縦・横の長さが長い方を 50pix の正方形になるように調整し，長さが短い方は中心に配置する（図 5）．

外接手型画像に対し，Depth 画像の各ピクセルの距離値から HOG 特徴量を計算することで手の大まかな形状を抽出する．本実験の HOG に関しては，1 セルあたり 5×5pix，1 ブロックあたり 3×3 セル，輝度の勾配方向を 0° ～359°，勾配方向数を 18 とした．このときブロックの移動回数が 64 回となるので，手型画像 1 枚あたりの次元数は 3×3×18×64 = 10368 次元である．しかし，HMM を使用するため

には次元数が非常に大きく，そのまま認識に用いることは困難であるため，得られた全指文字手型画像に対し，主成分分析を行うことにより次元数の削減をする．

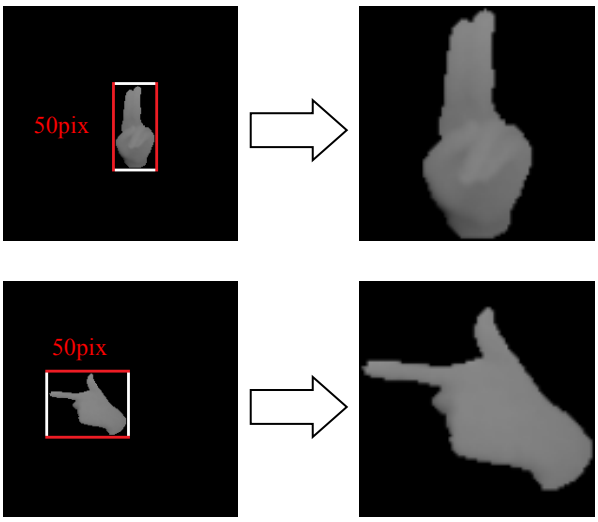


図 5　手型の外接矩形

3.3 HMM による連続認識

前述で得られた位置差分ベクトル⊿と主成分分析によって圧縮した HOG を用いて，HMM による連続認識を行う．本研究では Hidden Markov Model Toolkit(HTK)[13]に基づき，指文字単語の文字に対応する HMM モデルを作成し，特徴ベクトルからの出力確率が最も高くなるように学習を行う．そして未知の指文字が入力されたとき，すべての HMM で出力確率を計算して最も高い HMM に対応する指文字を認識結果とする．

本研究では位置差分ベクトル⊿と HOG における連続認識の精度を測るため，言語モデルは使用せず，文法フリーで行った．また，混合数は 1 とした．

4. 評価実験

4.1 実験条件

実験で使用する単語セットは，かな 46 文字が 2 回以上ずつ出現するように，また，固有名詞や外来語が数多く含まれるように選出した．単語の総数は 46，1 単語の文字数は 2 文字から 6 文字である．文字総数は 172 文字，1 単語の平均文字数は 3.8 文字，文字種類数は 62（濁音・半濁音・促音・拗音・長音含む）である．収録単語を表 1 に示す．

実験協力者は手話サークルに所属する聴者 4 名で，上述の指文字単語 46 個をそれぞれ 10 回ずつ収録して使用した．収録には Kinect を正面に 1 台と，指文字の確認用にビデオカメラ 1 台を用いて撮影を行った．また，腕には黒ベルトを巻いた状態で指文字を行った．

表 1　収録単語

01	あさい(アサイ)	24	ぱい(パイ)
02	あせろら(アセロラ)	25	はたはた(ハタハタ)
03	あてね(アテネ)	26	ぱなそにっく(Panasonic)
04	あめふらし(アメフラシ)	27	はのい(ハノイ)
05	あられ	28	ぱぱいや(パパイヤ)
06	うゆに(ウユニ[塩湖])	29	ばーみやん(バーミヤン)
07	えるされむ(エルサレム)	30	ひたち(日立)
08	おかか	31	ひのき(檜)
09	かかお(カカオ)	33	へちま(ヘチマ)
10	きゃんべら(キャンベラ)	34	へねしー(Hennessy)
11	けせんぬま(気仙沼)	35	ぺんね(ペンネ)
12	けにあ(ケニア)	36	ほととぎす(ホトトギス)
13	ここあ(ココア)	37	ほのるる(ホノルル)
14	じょなさん(ジョナサン)	38	まんごー(マンゴー)
15	しんがぽーる(シンガポール)	39	みりん(ミリン)
16	そふとばんく(SoftBank)	40	むえたい(ムエタイ)
17	つきじ(築地)	42	もすくわ(モスクワ)
19	でにーず(デニーズ)	43	ゆた(ユタ[州])
20	てりーぬ(テリーヌ)	44	よもぎ(ヨモギ)
22	とよた(TOYOTA)	45	ろーま(ローマ)
23	なます	46	わしんとん(ワシントン)

4.2 実験結果

本実験では実験条件を表 2，表 3 のように定義し，実験に使用した画像のフレーム数を表 4 に示す．なお，不特定手話者モデルー手話者クローズ，特定手話者モデルーデータオープンはクロスバリデーションとして Leave-one-out を 10 回行い，その平均を結果とする．また，文字単位の認識性能の評価には式 (1)の認識精度を用いる．ここで，全文字数とは，正解単語の総文字数を表す．

表 2　認識実験データー不特定手話者モデル

	学習データ	評価データ
手話者オープン	3 人×10 セット	1 人×10 セット
手話者クローズ	4 人× 9 セット	1 人× 1 セット

表 3　認識実験データー特定手話者モデル

	学習データ	評価データ
データオープン	1 人× 9 セット	1 人× 1 セット
データクローズ	1 人×10 セット	1 人×10 セット

表 4　フレーム数

実験参加者				合計
A	B	C	D	
27,976	29,178	27,126	32,156	11,6436

$$\text{認識精度}(\%) = \frac{\text{全文字数} - \text{置換誤り} - \text{挿入誤り} - \text{削除誤り}}{\text{全文字数}} \times 100 \quad (1)$$

予備実験として HOG 圧縮後 25，50，75，100 次元に位置差分⊿を加え，状態数を 3，5，7，9 と変化させ認識実験を行った．実験結果は不特定手話者モデルー手話者オープン 4 人の平均で，認識結果を表 5 に示す．

不特定手話者モデルー手話者オープン実験で最も認識精度の良かった 53 次元 9 状態について，その他の実験を行った結果を表 6，表 7 に示す．

表 5　手話者オープン結果（%）

	3 状態	5 状態	7 状態	9 状態
28 次元	34.0	48.1	53.0	53.6
53 次元	34.0	49.2	54.6	54.7
78 次元	34.0	49.2	52.8	53.2
103 次元	33.1	48.3	52.2	52.0

表 6　実験結果－不特定手話者モデル（%）

	実験参加者				平均
	A	B	C	D	
手話者オープン	61.5	55.5	46.7	55.2	54.7
手話者クローズ	88.1	82.0	87.0	92.5	87.4

表 7　実験結果－特定手話者モデル（%）

	実験参加者				平均
	A	B	C	D	
データオープン	93.5	90.8	93.3	94.5	93.0
データクローズ	95.4	94.7	95.3	96.3	95.4

4.3 考察

まず，特定手話者モデルのエラーについて分析を行う．エラーの内訳を表 8 に示す．「類似手型」によるエラーは指文字の手の型が似ていることによって生じるエラーで，「同一手型 異動作」によるエラーは，清音・濁音・半濁音の識別ができなかったものを指す．また，「同一文字 連結」は同じ文字が連続して行われたときに，一文字として認識されたものである．「その他」は上記で分類できなかったエラーを指す．

特定手話者モデルのデータクローズについて，非常に高い認識精度が得られることが一般的であるが，表 7 より 4.6%ほど下がることがわかる．最も多かったエラーが同じ文字が連続して行われる場合（「おかか」など），一文字と認識されてしまうエラー（「おか」など）で，全体のエラーの約 6 割を占めることがわかった．これは同一文字の連続のとき，手型の変化が小さく，文字を区切る腕の動きが不明確であるため自己遷移が大きくなってしまったからだと考えられる。本来同一文字の連続は腕を前方向に動かすスタンピングという動作をもとに，腕の動きの前後方向で区別しなければならないが，本実験では前後方向の特徴量は⊿z の 1 次元分の情報しか考慮されておらず，これらのエラーを改善するためには，この特徴に重みを付けるなどが考えられる．

次に多く見られたエラーが類似指文字によるエラーである．類似文字によって生じた置換誤りを表 9 に示す．また，図 6 に誤認識の多かった類似手型画像例を示す．圧縮した HOG 特徴量では指の本数などの細かい特徴まで把握しきれず，エラーが増えたと考えられる．また，「の」（人差し指でカタカナの「ノ」を書く動作）と「ー」（人差し指で縦に「ー」を書く動作）のように，手型と動きが似ている指文字の置換誤りも多く見られた．「ー」と「ん」（人差し指でカタカナの「ン」を書く動作）の置換誤りに関しては，「ん」の上から下へ下ろす動作の時が「ー」の動作と誤認識されたと考えられる．これらを改善するためには，指の本数が正確に反映されるような特徴量が望ましいと考えられる．動きを伴う指文字の場合は，文字の始点や終点の手型の情報を利用するなどが考えられる．

また，類似文字の削除誤りも多く見られた．これは上記二つのエラーが同時に起こり，ある単語の文字が前後の類似手型に誤認識され，さらにそれが前後の文字と同一文字と誤認識されてしまったと考えられる（例:「なます」→「ます」→「ます」）．この問題を解決するためには，上記同様にスタンピング動作を考慮した特徴量を考える必要があると考えられる．

その他のエラーに関しては，「ひのき」の「ひ」や「わしんとん」の「と」が削除されるエラーが見られた．これは前後に動きを伴う指文字がある場合，その動きのある指文字の渡りとして誤認識されたと考えられる．この問題を解決するためには，「同一文字 連結」と同様に，スタンピング動作から指文字のセグメントを正確に行う必要がある．

データオープンについて，表 8 からデータクローズのエラーに加え，類似手型による置換誤りが増えることがわかる．また，同一手型で動きの異なる指文字の誤認識（「は」を「ば」と誤認識など）も増えることがわかる．さらに，表 7 より実験参加者 A，C，D における認識精度がデータクローズに比べ，約 2%下がっているのに対し，実験協力者 B のみ 4%ほど下がっていることがわかる．このことから同一手話者の同一単語でもセットにより手型や腕の動きが安定していないことや，手話者間でもその安定度が異なることがわかる．これらを解決するためには，手型の安定

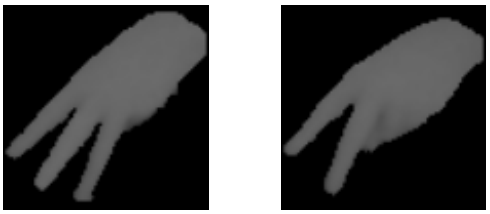
を図るために，同一手話者内での腕の動きの正規化や，手型の微妙な差を吸収できるような特徴量の検討が挙げられる．

表 8　特定手話者モデル エラー内訳（文字）

		データオープン	データクローズ
類似手型	置換	86	31
	挿入	2	0
	削除	55	54
同一手型 異動作		18	3
同一文字 連結		182	185
その他	置換	48	17
	挿入	21	0
	削除	65	24
合計		477	314

表 9　類似指文字例（文字）

指の数が減る文字		指の数が増える文字	
い→さ	1	す→ね	1
う→せ	3	に→み	1
す→ふ	1		
ま→な	7		
み→に	1		
る→れ	1		
型が似ている文字		型・動きが似ている文字	
す→ま	1	の→ー	1
も→た	1	ー→の	1
ろ→ら	3	ん→ー	7
		ー→ん	1



「ま」　「な」

「ろ」　「ら」

図 6　類似手型

次に不特定手話者モデルのエラーについて分析を行う．手話者クローズについて，表 6，表 7 から，特定手話者モデルーデータオープンより 5.6%認識精度が下がることがわかる．これは前述のエラーに加え，学習データの実験参加者数を増やすと，文字の学習モデルの分散が大きくなることから，より誤認識が増えると考えられる．同じ単語でも実験参加者により呈示方法が異なり，同じ場所でスタンピングを行う手話者や，左から右へずらしながらスタンピングを行う手話者が見られた．これにより，呈示する指文字の角度が異なり,認識誤りが増加したと考えられる.図 7 に呈示による角度の違いの例を示す．また，手の大きさや指の長さの違いも原因であると考えられる．類似手型に関しても，学習モデルの分散が大きくなることから，より誤認識が増える傾向がある．これらを改善するためには，角度や手の大きさの正規化を行う必要があると考えられる．

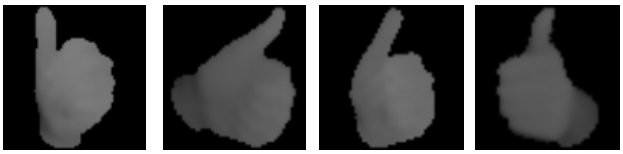


図 7　手話者間の呈示の違い例 「た」

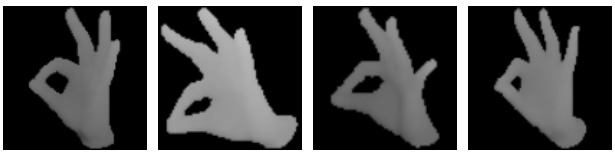
最後に手話者オープンについて，手話者クローズより 32.7%認識精度が下がることがわかる．手話者クローズ同様に手話者間での指文字の呈示の違いや，手や指の大きさの違いが大きく影響しているからであると考えられる．中でも一人だけ呈示方法が異なると別の類似文字に誤認識され，大きく認識精度が下がった．例として「ろーま」,「あめふらし」,「ここあ」が挙げられる．

「ろーま」に関しては，実験参加者 C のみ「ま」の呈示角度が異なり（図 8),「あめふらし」に関しては，実験参加者 B のみ「め」の呈示角度が異なり（図 9)，認識精度が下がった．これは HOG 特徴量が回転に弱いためであると考えられる．

「ここあ」に関しては，実験参加者 D のみ前述の同一文字の連結エラーに加え,「こ」を「ご」に認識するエラーが見られた．これは呈示するときに一人だけ左から右へずらしながら呈示したため，濁音と誤認識されたからであると考えられる（図 10)．このことから，腕の動きにも手話者での個人差が大きいということが考えられる．これらを改善するためには，HOG 特徴量は回転に弱いため，回転を考慮した個人差を小さくできるような特徴量の検討や，呈示する角度の正規化などが考えられる．

A　　B　　C　　D

図 8　手話者間の呈示（角度）の違い例　「ま」



A　　B　　C　　D

図 9　手話者間の呈示（角度）の違い例　「め」

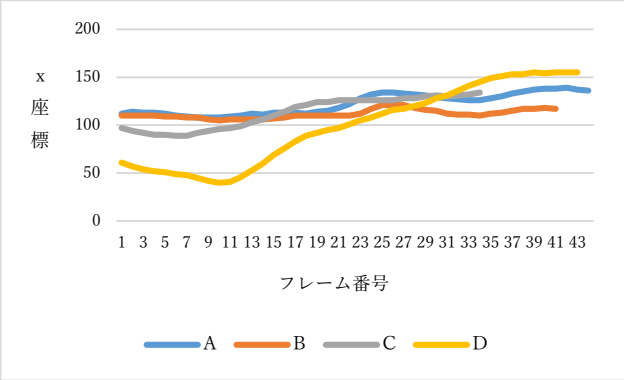


図 10　手話者間の呈示（動き）の違い例　「こ」

以上より，指文字には手の向きや角度，呈示する際の腕の動きなど個人差が大きいことがわかった. また表 6, 表 7 より手話者が特定されているモデルがより高い認識精度を持つことがわかる．このことから，個人差の問題を考慮した認識手法としてマルチテンプレートなどの手法が挙げられる．

5. おわりに

本研究では連続で指文字認識を行うため, HOG 特徴量を用いた HMM による連続認識手法を提案し，評価実験を行った．その結果，特定手話者モデルーデータクローズでは 95.4%，不特定手話者モデルー手話者オープンでは 54.7%の認識精度を得た.

今後の課題として，本実験で識別が困難であった類似した手型や同じ文字が連続する場合の認識精度を向上させる手法の検討が挙げられる. そのためには手型特徴量の再考，あるいは別の特徴量との組み合わせなどが考えられる．さらに, スタンピングをより考慮した認識手法の検討も行う.

また，今回の実験協力者は 4 人と少人数であったため，手型のサンプル数が少なく，話者間での差が大きくなってしまった．よって実験協力者の数を増やし，個人差の影響を小さくすることが挙げられる. そのため実験協力者の数，データ数を増やした実験を行う必要があると考えられる.

謝辞　本研究を行うにあたり，実験にご協力頂いた手話サークルの方々に深く感謝いたします．また，本研究は文部科学省科学研究費補助金基盤研究(C)15K00223 の補助を受けています.

参考文献

[1] 後藤岳志, “動作を伴う指文字および連続した指文字認識システム”，北陸先端科学技術大学院大学修士論文，1993
[2] 江本祐太，宮島千代美，江藤克亘，武田一哉，“HMM に基づく連続指文字認識・合成用コーパスの構築”，電子情報通信学会，IEICE Tecnicsl Report，SIP2005-94，2005
[3] 舟川政博，平山亮，“指文字画像からの手指形状特徴量抽出方法の検討”, FIT2006(第 5 回情報科学技術フォーラム), l_037, pp.87-88，2006
[4] 浜田康弘，島田伸敬，白井良明，“遷移ネットワークに基づく多視点画像時系列からの手指形状推定”, 電子情報通信学会論文誌，Vol.J85-D-II，No.8，pp.1291-1299，2002
[5] 菊田智也，“類似文字を考慮した段階的な指文字認識”，三重大学大学院工学研究科修士論文，2011
[6] K.Otiniano-Rodriguez，G.Camara-Chavez，“Finger Spelling Recognition from RGB-D Information using Kernel Descriptor”，2013 XXVI Conference on Graphics,Patterns and Images，pp.1-7，2013
[7] 三浦航平，張英夏，向井信彦，“爪と手首の位置検出に基づく日本語手話の指文字認識”，映像情報メディア学会技術報告，ITE Technical Report Vol.37，No.17，pp199-202，2013
[8] 織茂裕介，玉國祐司，高橋大介，岡本教佳，“Kinect を用いた指文字認識の検討”，映像情報メディア学会技術報告，ITE Technical Report Vol.38，No.9，pp.31-32，2014
[9] Kuan-Yu Chou，Hui-Chi Chuang，Meng-Tzu Chiu and Yon-Ping Chen，“Real-Time Fingerspelling Recognition System Design Based on RGB-D Image Information”，2014 CACS International Automatic Control Conference，pp.75-80，2014
[10] Akira Nagasue，Joo Kooi Tan，“Japanese Finger-spelling Recognition Using a Chest -mounted Camera”，SICE Annual Conference 2012，August 20-23，2012
[11] 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技術科学研究科修士論文，2012
[12] N.Datal，B.Triggs，“Histgrams of oriented gradients for human detection”，Proc.IEEE Conference on Computet Vision and Pattern Recognition (CVPR)，pp.886-893，2005
[13] S. Young，et al;The HTK Book (for Version 2.2)