距離画像を用いた動きをともなう指文字認識に関する基礎的検討

筑波技術大学大学院　技術科学研究科　産業技術学専攻 [1)]
筑波技術大学　産業技術学部　産業情報学科 [2)]

三宅太一 [1)]　若月大輔 [2)]　内藤一郎 [2)]

要旨：音声認識とその関連技術の進歩によって、音声で様々な操作を容易に入力できる情報端末が普及してきた。しかし、聴覚障害者にとっては音声の発音が難しく、音声入力をうまく利用できないことが多い。そこで、音声の代わりに聴覚障害者が日常のコミュニケーションで使用している指文字による入力を実現するために、距離画像を用いて動きをともなうすべての指文字を認識する方法について基礎的な検討を行った。本報告では距離画像を用いた手領域の抽出、および手型とその動きを認識する手法について述べ、手型が静止した指文字と濁音、半濁音の文字、および拗音に用いる小さい文字を含めた指文字の計 75 文字を認識させた結果について報告する。

キーワード：指文字認識、距離画像、赤外線 TOF カメラ、聴覚障害者、入力インタフェース

1. はじめに

音声認識とその関連技術について研究開発が進み、音声入力によって様々な操作が可能な情報端末が発売され普及してきた。しかし、聴覚障害者のなかには発音が困難で不明瞭な人が多く、音声入力機能をうまく利用できない場合がある。聴覚障害者の発音に特化した音声認識の学習データの構築も考えられるが、発音の個人差が大きく高い認識率を得るための調整が困難である。そこで本研究では、聴覚障害者が日常的に使用している指文字を活用した入力インタフェースの実現を目指し、動きをともなうすべての指文字を認識する方法について研究を進めている。

指文字には図 1 に示すように各清音に対応した手型があり、このうち静止した手型で表現するものが 41 文字、手型そのものを動的に変化させて表現するものが 5 文字ある。また、濁音、半濁音の文字、拗音と促音に用いる小さい文字（以後、小書き文字と呼ぶ）と長音の文字については、清音の手型を維持したまま図 2 のように決まった方向に平行移動させて表現する。これらの指文字は合計で 36 文字あり、そのすべてを認識する必要がある。

本報告では、赤外線 TOF 方式の距離画像カメラを用いて動きをともなうすべての指文字を認識するための基礎的な検討を行った。指文字を撮影した距離画像から手型部分の領域を抽出する方法、濁音、半濁音および小書き文字のように手型を変化させずに平行移動させて表現する動きをともなう指文字の認識方法を提案する。また、実例として手型が静止した指文字 41 文字と、動きをともなう指文字 34 文字（本報告では促音と長音の 2 文字は除外した）の計 75 文字について認識した結果について報告する。

図 1　清音指文字の表現

図 2　動きをともなう指文字の表現

2. 指文字認識の関連研究

指文字認識の関連研究では、手型が静止した指文字を対象とした報告が多く、動きをともなう指文字を含めた指

文字の認識を試みた例は少ない。

手型が静止した清音の指文字を認識する関連研究の方法は、指文字を撮影した輝度画像（モノクロ、カラー画像などの各画素に輝度値を格納した画像）や距離画像（各画素に距離値を格納した画像）に対して画像処理を行い非接触で認識を行う方法と、データグローブなどの接触式のセンサを用いて得られるデータをもとに認識する方法の 2 種類に大別できる。

前者の画像処理によって指文字の認識を行う方法では、輝度画像を 2 値化して手の領域を抽出し、手型を認識する方法が提案されている [1]。また、手領域をより認識率良く抽出するためにカラーグローブを装着した手を撮影してカラー情報から手領域の抽出を行い、認識を試みた方法も提案されている [2]。しかし、これらの方法の多くは照明や背景の条件がそろっていなければ、手領域のロバストな抽出や手型の認識率を向上させることができない。

一方、後者のデータグローブを用いて各指の屈伸などの情報を取得して認識する方法では、アルファベット文字を高い認識率で識別できることが報告されている [3]。しかし、これらの接触式の方法の多くは装用者への負担は少なくなく、手型の動作範囲が制限されてしまう場合があり、自然な指文字表現を妨げる可能性がある。

3D スキャナを用いて取得した距離画像を用いて画像処理で認識する方法では、清音の静止指文字を識別することが可能である [4]。また、距離画像では各画素の距離によって背景と手の領域を容易に分離しやすく、認識対象の指文字を抽出しやすい。しかし、3D スキャナは解像度と認識率が高い距離画像を取得できる反面、フレームレートが著しく低いため動きを含む指文字への対応が難しい。一方、毎秒 30fps 程度で距離画像を取得できる PrimeSense 社の Light Coding 方式を採用したセンサ（Microsoft 社、Kinect）を用いて指文字を認識し、入力インタフェースを試作した研究 [5] も報告されているが、動きを含んだ指文字は認識対象としていない。

本研究では先行研究の課題を解決し、動きをともなう指文字を含めたすべての指文字の認識を行うために、赤外線 TOF カメラ [6-7] で撮影した距離画像を用いた認識を試みた。同カメラは撮影対象に赤外線を照射し、対象で反射されて戻って来るまでの時間から距離情報を計測する。赤外線光は目に見えず、データグローブ等の接触式センサを用いた計測ではないため、利用者に対する物理的、精神的な負担は少ない。

本報告では、図 1 で示した清音指文字 46 文字のうち、手型を変化させて表現する「の」、「も」、「り」、「を」、「ん」を除く手型が変化しない静止指文字の 41 文字と、濁音、半濁音、および拗音を表現する小書き文字のように、手型を平行移動させて表現する動きをともなう指文字 34 文字の、合計 75 文字を対象とした認識方法について述べ、実験を行った結果について報告する。

3. 動きをともなう指文字の認識

3.1 認識処理の概要

距離画像を用いた動きをともなう指文字の認識の流れ a ～ d を次に示す。

a. 指文字の距離画像の撮影
b. 距離画像から手領域の抽出
c. 抽出された手領域の手型の認識
d. 手型の動きの認識

まず、処理 a で赤外線 TOF カメラを用いて対象となる指文字の距離画像を撮影し、処理 b でその距離画像から手領域のみを動的に抽出する。次に、抽出された手領域の手型を処理 c で認識を行い、処理 d でその手型の動きを認識する。動きをともなう指文字については、処理 c と d の結果を組み合わせて認識を行う。たとえば、「か」の手型で動作がなければ「か」、横方向に動作があれば濁音の「が」として認識する。

3.2 指文字の距離画像の撮影

距離画像とは各画素に距離値を格納した画像で、撮影には赤外線 TOF カメラの SR-3000（Swiss Ranger 社）を用いた（図 3(a)）。同カメラの距離画像の解像度は 176 × 144、最大フレームレートは 50fps である。図 3(b) は距離が近い画素を明るい、遠い画素を暗い輝度で表示した結果である。各画素の距離値をカメラパラメータをもとに逆射影変換することで、図 3(c) のように各画素の 3D 頂点座標値を得ることができる。

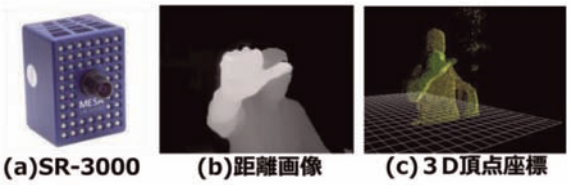

図 3　SR-3000 の動作イメージ

3.3 距離画像から手領域の抽出

距離画像カメラに対して一番近い距離にある物体を手領域と仮定して抽出を行う。本研究では距離画像を用いて撮影された物体の体積を手がかりにして手領域を抽出する。図 4 に手を撮影した様子を示す。距離画像中のある i 番目の画素は、撮影した距離画像カメラからの距離 d に対して四角錐台を形成する。この四角錐台の体積 $V(i,d)$ は、

$$V(i,d) = s_i * d - s_i * d_i \quad \ldots (1)$$

で計算できる。ただし、s_i は i 番目の画素が占める単位距離あたりの面積、d_i は i 番目の画素の距離値である。

まず、各画素をその距離値が小さい順にソートを行い、ソートした j 番目の画素について式 (1) で四角錐台の体積を計算できるようにする。次に、距離値が小さい順に n 個の画素が形成する四角錐台群の体積の合計 $V(n)$ を、

$$V(n) = \sum_{j=0}^{n-1} V(j, d_{n-1}) \quad \cdots (2)$$

で計算する。

本研究では式 (2) で計算される手領域の体積 $V(n)$ は、実際の手の体積の半分程度であり、手型によらず一定であると仮定する。つまり、体積のしきい値 V_t を設け、n を 1 から順に大きくして $V(n)$ を計算し $V(n) \geqq V_t$ となる n を求める。最後に、j=0,1,...,n-1 に対応する画素を手領域として抽出する。

抽出された領域を矩形領域で切り出して16×16にスケーリングする。各画素の距離値を特徴量として 256 次元の特徴ベクトルを次節で述べる手型の認識に使用する。

3.4 抽出された手領域の手型の認識

手型の認識は k 近傍法 [8] で行う。手型の特徴ベクトル x とその手型クラス C の組 (x,C) を 1 つの学習データとし、N 個の学習データ (x_i,C_p) を k 近傍法の学習データとして登録する。i (i=0,...,N-1) は学習データのインデックス、P は静止指文字のクラスの種類を表すインデックスである。例えば、C_0 は「あ」、C_1 は「い」のクラスを表す。

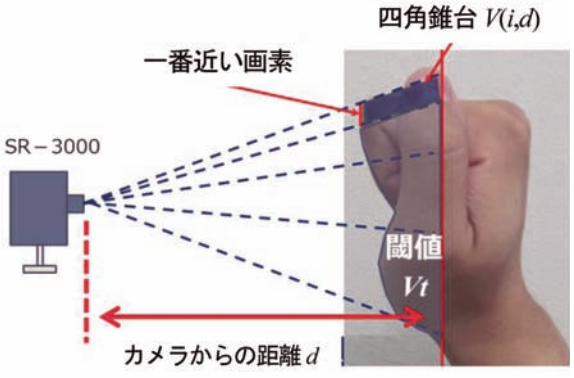

図 4　手型抽出イメージ

今、入力データとして x_{input} が与えられた時、x_{input} とすべての学習データ x_i との間のユークリッド距離 D_i を計算して、D_i が小さい順にソートし、先頭から k 番目までの C_p について出現数が最も多いクラスを認識結果として出力する。

3.5 手型の動きの認識

本研究では、手型の移動方向、および手型の移動距離を求めることで、濁音、半濁音に対応する指文字、ならびに小書き文字に対応する指文字の認識を行う。

手型の移動方向を調べるために、連続した任意のフレーム数 N_f を決めておき、各フレームの手型の重心（3D 頂点座標群の重心）をまとめた重心群に対して主成分分析（PCA）をかける。PCA により得られた重心群の分散が最も大きい方向、つまり第一主成分方向を手型の移動方向とする。

3D 空間上の x、y、z 軸方向をそれぞれ濁音、半濁音、拗音に用いる小書き文字の手型の動作方向を表す軸 u_i (i=0,1,2) であると図 5(a) のように定義し、PCA の第一主成分方向の単位ベクトル v とこれらの軸とのなす角が最小になる軸 u' を選択する。ここで選択された軸を手型の動作方向の候補とする。 つまり、

$$max\ \{abs(\boldsymbol{v} \cdot \boldsymbol{u_i})\} \quad \cdots (3)$$

を満たす u_i を u' とする。

この u' を用いて手型の移動距離である L を求める。図 5(b) のように第一主成分のスコアの最大値を S_{max}、最小値を S_{min} とすると、移動量 L は

$$L = (S_{max} - S_{min})\boldsymbol{v} \cdot \boldsymbol{u}' \quad \cdots (4)$$

で計算される。この時の移動距離 L が、移動距離のしきい値 L_t に満たない場合は、表現している手型が静止していると判別する。

4. 指文字の認識実験

4.1 実験環境と条件

図 1 で示した清音指文字のうち手型を静止して表現する指文字 41 文字と、手型を平行移動させて指文字を表現する濁音、半濁音ならびに拗音に用いる小書き文字の指文字 34 文字の計 75 文字を対象として認識実験を行った。被験者 1 名が提示した指文字を各クラス 10 枚ずつ保存（計 410 枚）したものを k 近傍法の学習データとして使用した。学習データを取得した人と同一者が 1 クラスにつき 20 回の試行を行い、その認識率を調べた。式 (3) (4) の各パラメータの設定は、k 近傍法の k=7, 保存した重心のフレーム数 N_f=5, 移動距離のしきい値 L_t=0.1m とし、カメラからの距離 0.7m の位置で手型を提示した。

4.2 実験結果

4.1 節で述べた実験環境条件で認識を行った結果を、

表 1　認識実験結果（清：清音，濁：濁音，半濁：半濁音，小：小書き文字）

クラス	清	動きをともなう		
		濁	半濁	小
あ	1.00	−	−	1.00
い	0.40※	−	−	0.00※
う	1.00	−	−	1.00
え	1.00	−	−	1.00
お	1.00	−	−	1.00
か	1.00	1.00	−	−
き	1.00	1.00	−	−
く	1.00	1.00	−	−
け	1.00	0.80	−	−
こ	0.80	1.00	−	−
さ	1.00	1.00	−	−
し	1.00	0.50※	−	−
す	0.60※	0.40※	−	−
せ	1.00	1.00	−	−

クラス	清	動きをともなう		
		濁	半濁	小
そ	0.00	0.80	−	−
た	0.80	0.80	−	−
ち	1.00	0.50※	−	−
つ	0.85	0.40※	−	−
て	1.00	1.00	−	−
と	0.80	1.00	−	−
な	1.00	−	−	−
に	1.00	−	−	−
ぬ	0.20	−	−	−
ね	1.00	−	−	−
は	1.00	0.40※	0.40※	−
ひ	0.35※	0.60※	0.80	−
ふ	1.00	1.00	1.00	−
へ	1.00	1.00	1.00	−

クラス	清	動きをともなう		
		濁	半濁	小
ほ	1.00	1.00	1.00	−
ま	0.40※	−	−	−
み	0.85	−	−	−
む	0.90	−	−	−
め	1.00	−	−	−
や	1.00	−	−	1.00
ゆ	1.00	−	−	1.00
よ	1.00	−	−	0.80
ら	0.65※	−	−	−
る	1.00	−	−	−
れ	1.00	−	−	−
ろ	0.85	−	−	−
わ	1.00	−	−	1.00
Avg	0.86	0.81	0.84	0.87
全体	**0.85**			

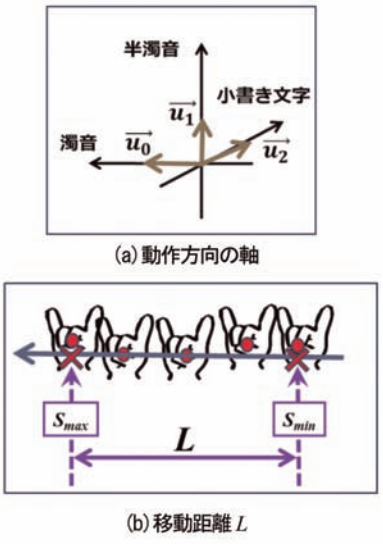

図 5　手型の移動距離の計算

表 1 に示す。表中の※印がついている数値は、認識率が全体の認識平均よりも低かったものを示す。対象とした指文字全体の平均認識率は約 0.85 となった。

5. 考察

5.1　静止指文字の誤認識

表 1 の結果のうち、特に認識率が低かったのは類似した手型の指文字であった。具体例を図 6 に示す。「ひ」と「ら」については相互に誤認識し合ってしまい認識率が低かった（図 6(a)）。「い」、「そ」、「ぬ」、「ま」については、それぞれ「ち」、「は」、「ろ」、「ね」に誤認識され、認識率が低かった（図 6(b)）。

類似した手型について誤認識が生じる原因として、識別器の識別能力の限界、学習データ不足、特徴ベクトルの限界が挙げられる。本報告で使用した識別器は k 近傍法であり、最も単純で基本的な線形識別器である。似ている手型の特徴ベクトルに対して線形識別器では十分に識別できなかったものと考えられる。学習データについても各 クラス 10 個ずつでは少なかったものと考えられる。今後は、非線形識別器を導入し、識別器に合わせた学習データ数について検討する必要がある。また、本報告で使用した特徴ベクトルは、16 × 16 にスケーリングした距離画像の各画素をそのまま用いた 256 次元ベクトルである。あまり特徴を記述していない成分が含まれており、冗長であった可能性もある。そこで、PCA で最適化した特徴ベクトルについて追加実験を行い、その認識率について調査を行った。

5.2　特徴ベクトルの最適化による認識率の変化

元の 256 次元の特徴ベクトル PCA を用いて 20 次元まで次元削減を行った。特に認識率の低かった指文字を中心に全く同じ条件で認識率を求めた。その結果を表 2 に示す。特徴ベクトルの次元削減を行う前では、認識率の低かった静止指文字の平均認識率は 0.37 だったが、次元削減後では 0.64 となり、認識率の向上が確認できた。 しかし、次元削減を行った後でも「ぬ」の指文字ついては認識率の改善は見られなかった。追加実験の結果から、次元削減によって全体の平均認識率の向上が見られたが、一部のクラスの認識率は改善されないことが分かった。

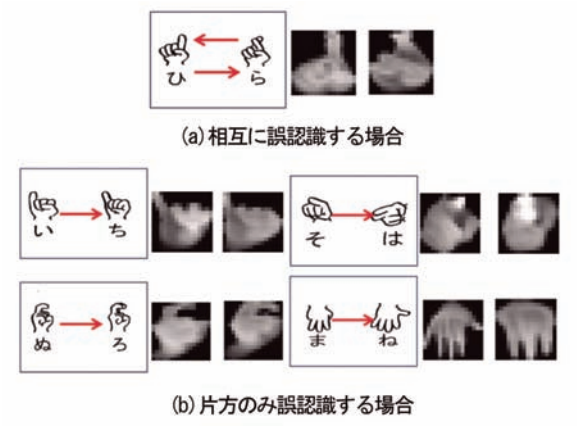

図 6　誤認識の具体例

表 2　PCA による次元削減後の認識実験結果

クラス	PCA (256次元)	PCA (20次元)
い	0.40	0.8
す	0.60	0.75
そ	0.00	0.50
ぬ	0.20	0.00
ひ	0.35	0.70
ま	0.40	1.00
ら	0.65	0.70
平均認識率	0.37	0.64

5.3　動きをともなう指文字の誤認識

動きをともなう指文字を誤認識してしまうパターンは、移動距離が足りず静止指文字として認識されてしまう場合と、動作中の手型が誤認識されてしまう場合の 2 つであった。前者は、手型の移動距離のしきい値 L_t の調整不足が原因であると考えられる。 後者は、手型の動作中に各フレームの距離画像が移動方向に対して残像することが原因である。例えば「び」は、人差し指 1 本を立てた「ひ」の手型を横方向に動作させるが、人差し指と中指を立てる「う」や「ら」に誤認識されることがあった。

これらの誤認識の解決策として、動きをともなう指文字について、手型の移動速度、移動距離、およびその間の連続的な手型の距離画像を学習データとした特徴ベクトルを設け、パターン認識により識別させる方法が考えられる。

6. まとめ

本報告では、指文字を入力インタフェースとして活用するための基礎的な検討として、動きをともなう指文字を含めたすべての指文字を認識する方法を提案した。提案法では、距離画像を用いて手型を動的に抽出して識別を行い、PCA で手型の動きを判別することで指文字を認識する。

実例として清音の指文字、濁音と半濁音の指文字、および拗音に用いる小書き文字の指文字を認識させた結果、認識率の平均値は約 0.85 となった。

今後は、実験で明らかとなった課題を解決するために、画像全体の相関を 35 次元の特徴量として表現した高次局所自己相関特徴（High-order Local Auto-Correlation :HLAC）[9] や、HLAC を時間軸方向に拡張し動画に適用できるようにした立体高次局所自己相関特徴（Cubic High-order Local Auto-Correlation:CHLAC）[10] を特徴ベクトルとして用いることで、類似形状や動きをともなう指文字に対する認識率の向上を目指す。

また、認識率が高い識別器とされるサポートベクターマシン（SVM）[11] や AdaBoost[12] を実装して認識率を評価し、識別器の違いや学習方法の違いで認識率にどのような差が出るのかも含めて検討を進めていきたい。

参考文献

[1] 諸角建：画像情報を利用した指文字の認識. 拓殖大学理工学研究報告 9(2)：pp.51-60, 2004.

[2] 新澤真郷, 大矢誠：画像処理による指文字の認識. 日本機械学会第 46 期総会･講演会 講演論文集 97(1)：pp.419-420, 2008.

[3] 福島大志, 宮崎文夫, 西川敦:指文字入力インタフェース「Fingual」の開発. インタラクション 2011 論文集：2011.

[4] 王宇, 板井聖治, 小野智司：PCA と 3D スキャナによる指文字認識. 情報知識学会誌 16(1)：pp.51-60, 2004.

[5] Pugeault. N,and Bowden. R：Spelling it Out:Real-Time ASL Fingerspelling Recognition. 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, Jointly with ICCV 2011：pp.1114-1119, 2011.

[6] T. -Ogger,, M. Lehmann, R. Kaufmannn, M. Richter, P. Metzle, G. Lang, F. Lustenbergr, N. Blanc, An all-solid-state optical range camera for 3D- real-time imaging with sub-centimeter depth-resolution (SwissRanger), Proc. SPIE Vol.5249, pp. 634-545, 2003.

[7] Buttgen, B., Oggier, T., Lehmann, M.: CCD/CMOS Lock-in pixel for range imaging:challenges, limitations and state-of-the-art. In proceedings: 1st range imaging research day, pp.21-32, Ingensand / Kahmann (eds.), Zurich, 2005.

[8] 雨宮秀文, 佐野健吾：k-NN 法による学習を用いた感性語による画像検索全国大会講演論文集 第 56 回平成 10 年前期(3), pp.142-143, 1998.
[9] 小林匠 , 大津 展之：画像特徴量 – 高次局所自己相関に着目した画像特徴量と画像認識への応用 , 電子情報通信学会誌, 94(4), pp. 335-340, 2011.
[10] 南里卓也, 大津展之：複数人動画像からの異常動作検出.情報処理学会論文誌 コンピュータビジョンとイメージメディア 45 (SIG_15), pp.43-50, 2005.
[11] Nello Cristianini, John Share-Taylor 著, 大北剛 訳: サポートベクターマシン入門. 共立出版, 2005 年
[12] 三田雄志：AdaBoost の基本原理と顔検出への応用：CVIM 研究会 チュートリアルシリーズ (チュートリアル 2). 情報処理学会研究報告, CVIM.2007(42), pp. 265-272, 2007.

A Basic Study on Recognizing Fingerspelling with Hand Movements By the Use of Depth Image

MIYAKE Taichi[1)], WAKATSUKI Daisuke[2)], NAITO Ichiro[2)]

[1)]Graduate School of Technology and Science, Tsukuba University of Technology
[2)]Faculty of Industrial Technology, Tsukuba University of Technology

Abstract: IT devices that can easily be operated in a speech input have been popularized by the improvement of speech recognition and related works. However, hearing impaired persons often use speech input because they may have difficulty speaking clearly. Therefore, to achieve fingerspelling input rather than speech input, we studied fingerspelling recognition. This paper describes a method to recognize Japanese fingerspelling based on hand movements selected from a depth image. Our method recognizes the shape and movement of hand area selected from a depth image. We report the results of an experiment conducted to recognize 75 characters of Japanese fingerspelling that included voiced, semi-voiced, contracted, assimilated, and prolonged sounds.

Keywords: Fingerspelling recognition, Depth image, 3D time-of-flight camera, Hearing impaired person, Input interface