画像処理による手話単語認識

Recognition of sign language words by image processing

○鳥毛　明(成蹊大)　内田　和貴(成蹊大)
Akira TORIGE, Seikei University
Kazuki UCHIDA, Seikei University

Key Words: Sign Language, Image Processing, Welfare Engineering

1.　序論

1.1　研究背景

現在、様々な企業や大学等の研究機関で手話の機械翻訳の研究が行われており、主にカメラ、距離画像センサー、身体に装着した動作検出用の電極等が用いられている（Fig.1 参照）。

Fig.1　Sign language translation that uses camera and cyber glove by Hitachi, Ltd. (2001)

カメラの他にセンサを用いることで顔の表情だけでなく手指の動きをより正確に計測することができるが、その分大掛かりな装置が必要になる他、センサを身体に装着して手話を行うという行為は手話操者の精神的な負担にもなるため、より簡潔な仕組みで手話通訳できるようにすることが求められる。

そこで、本研究では手話操者の身体にセンサ等を装着したり特殊な服装などを着せたりせず、なおかつ一般的な服装でカメラのみを用いて手話通訳を行う方法を考案した。

1.2　研究目的

本研究では以下の２つを研究目的とした。

1．手話操者への負担を減らすための新たな手話翻訳システムの開発。
2．開発したシステムの翻訳精度検証実験。

2.　動作認識システム

2.1　システム概要

本研究では汎用的な環境で手話動作を認識するため、特定の撮影環境や手話操者の服装に依存せず、なおかつ手話操者になんらかのセンサー等を装着することなく手話動作を認識することを目標とした。そこで、動作認識の手段として色検出の代わりにモーション検出処理を採用した。色検出を使わないことにより特定の撮影背景や服装に依存せず、さらに手の動きをカメラ映像のみで認識するため手袋を等を装着する必要がなくなった。

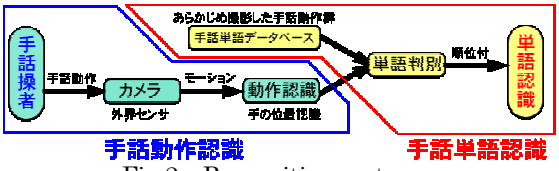


Fig.2　Recognition system

2.2　撮影環境

本研究で構築したシステム概要を Fig.2 に撮影環境を Fig.3 に示し、Table 1 に使用した機器の仕様を示す。PC に USB カメラを接続して手話操者を撮影し、撮影された映像から手話単語を判別する仕組みとなっている。また、確認用モニターに撮影映像がそのまま表示されており、手話操者はモニターを見ることで自分の立ち位置や撮影領域を確認できるようになっている。暗幕や手袋を用いた環境設定や、カメラ以外のセンサは用いない。

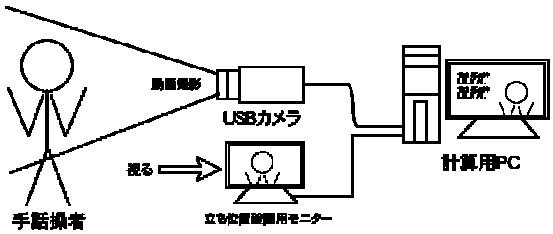


Fig.3 Word recognition taking a picture environment

Table.1 Taking a picture environment

PC	CPU	Core i7 920
	OS	WindowsXP
	image library	Intel OpenCV
Camera	Video camera	Xacti DMX-HD2000
	USB camera	CMOS130-USB2

3.　単語辞書

人間が言語を使用して相手と会話するためは俗に言う「語彙」が必要となる。手話においても同様で、単語別に共通認識となる動作が「語彙」に相当する。本研究では人工的な「語彙」を再現するために、複数人の手話動作を撮影して単語辞書を作成した。Fig.4 に単語辞書を作成するための環境を示す。

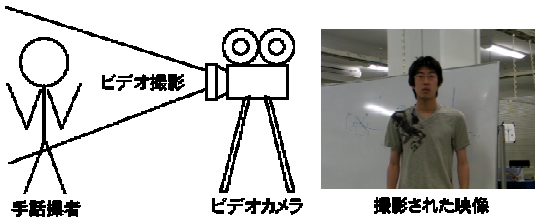


Fig.4 Word dictionary making environment

4　動作認識システム

4.1　動作認識システム概要

本研究で開発した動作認識システムは映像から手話操者の両手の位置をそれぞれ認識する。手話は基本的に手腕しか動かさないため、背景が動かないという前提であれば、映像内での動体は必ず手腕であると考えることができる。また、両

手座標を手話操者の顔を基準とすることで、手話操者の身長や映像内の立ち位置が違っても対応できる（Fig.5 参照）。

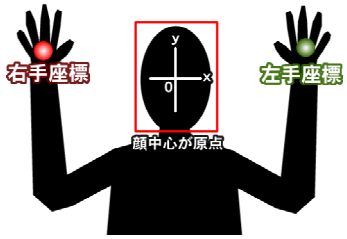


Fig. 5　Both hands coordinate

4.2　動作認識プロセス

まず、手話操者がカメラ映像内の中央付近に位置するようにカメラの位置と向き、手話操者の立ち位置を決定する。単語によっては手話動作中に手が画面外に出ないようにカメラの向きを調整することも必要になる。撮影した映像は USB ケーブルを通して PC に転送される。

次に、撮影背景から物体（本研究では手話操者）だけを抽出する。動いている物だけを抽出するので手話操者が撮影中にしばらく動かない状態でいると背景と同化する。Fig.6 に撮影画像(右)と物体抽出後の画像(左)を示す。

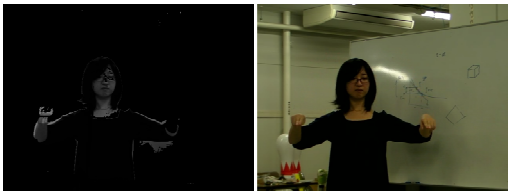


Fig. 6　Object extraction by dynamic background update

それから、オプティカルフローの一種であるブロックマッチング法を用いて画像内から動きを複数のベクトルとして抽出する。本研究ではベクトル方向の値は使用せず、ベクトルが作られている特徴点のみを使用する。Fig. 7 に物体抽出後の画像(右)とオプティカルフローを赤線で表した画像(左)を示す。

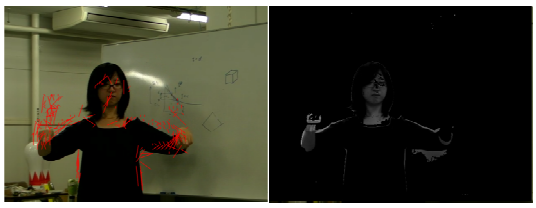


Fig. 7　Block match method

この時点で映像から動作を抽出できたが、各ベクトルが右手と左手どちらの動作なのか判別する必要がある。そこで、クラスタリングという手法を用いて動作ベクトル特徴点の集合を２つのグループに分ける。さらに、手が動いていれば手の周りに特徴点が集まるため、それぞれのクラスタの重心座標を求めて両手座標候補点をとしている(Fig.8 参照)。

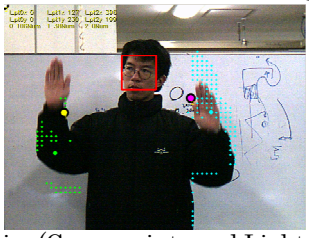


Fig. 8　Clustering(Green points and Light blue points)

本研究の両手座標は顔座標を基準としているため、顔認識で手話操者の顔座標を得る必要がある。Fig. 9 の赤矩形線は顔認識アルゴリズムによって顔の座標が分かる様子を表している。

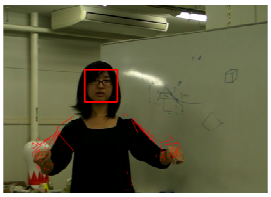


Fig. 9　Facial recognition

最後に、両手座標候補点と顔座標から得られる推定両肩座標から仮の両手座標を求め、さらに前フレーム以前の両手座標を組み合わせてＦＩＲフィルターを通すことで最終的な両手座標を決定する。片腕しか使わない手話の場合は両手座標候補点が写っている片腕に集まるため、片腕だけの場合でも強引に両手座標を計算する（Fig.10 参照）。

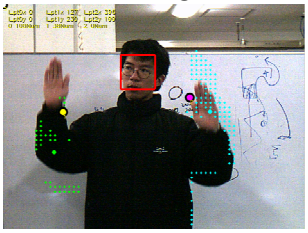


Fig. 10　Both hands decision
(yellow point and Purple point)

5.　手話単語認識システム

手話単語を比較して一致評価を出す方法について説明する。ここで言う一致評価とは、撮影された動作に対してどの程度単語の動きが似ているかどうかを数値で表したものである。単語を比較認識するには、「あらかじめ作成した単語辞書データ」と「撮影データ」を比較して差異を計算する必要がある。認識手順としては、

①　あらかじめ全単語の辞書データを作成
②　本番として手話撮影を行いつつ辞書データと比較
③　差異の合計が最も小さい単語ほどを正解であると認識

となる（Fig.11 参照）。

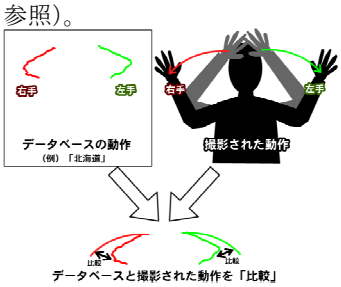


Fig. 11　Motion comparison

5.1　比較パラメータ

手の動作を数学的に比較するためには様々な比較要素が考えられるが、本研究では「静的位置」と「時系列動作」を比較要素として考えた。静的位置としては手話の「開始座標」、「平均座標」、「終了座標」を、時系列動作としては「フーリエ級数係数」をパラメータとして用いることにした。なお、各比較パラメータは手話操者の顔中心を座標基準とした右手 XY 座標と左手 XY 座標を基にして算出される。

１、静的位置パラメータ

右手と左手の開始座標、平均座標、終了座標を比較パラメ

ータとする。左右の手、XY 座標、開始平均終了を選択すると全部で 12 パラメータになる。
静的位置パラメータ ＝ 2×2×3 ＝ 12 パラメータ

２、時系列動作パラメータ
右手と左手の動作(座標変化の様子)をフーリエ級数係数（１～２次）を用いて表す。左右の手、XY 座標、正弦、余弦級数係数、１～２次係数を選択すると全部で 16 パラメータになる。
時系列動作パラメータ ＝ 2×2×2×2 ＝ 16 パラメータ

上記の「静的位置」と「時系列動作」のパラメータを合わせると全部で 28 パラメータになり、これが本研究で使用する全比較パラメータである。
静的位置+時系列動作 ＝ 12＋16 ＝ 28 パラメータ

３、時系列動作パラメータの計算式
時系列動作パラメータの計算式は手話開始から終了まで計算すると以下のように算出される。

フーリエ正弦級数係数パラメータ： $\frac{1}{N}\sum_{t=1}^{N} g_t \times \sin\frac{2M\pi t}{N}$

フーリエ余弦級数係数パラメータ： $\frac{1}{N}\sum_{t=1}^{N} g_t \times \cos\frac{2M\pi t}{N}$

N	全コマ数
gt	時系列XY座標値
t	コマ数(0～N)
M	フーリエ係数(1次、2次)

Exp. 1　Fourier series parameter

5.2　重み付け

人によって手話動作が違っても相手に意志を伝えることができるのは、動作の中に共通の要素があるためである。本研究ではこの性質を実現するため、個人差が小さい要素を重視、個人差が大きい要素を軽視することで人間の持つ共通理解の性質を模倣している。具体的には、要素パラメータごとの個人差のバラツキ具合（分散）を「重み」とし、カメラで撮影された手話単語の各要素パラメータにそれぞれの「重み」を掛け合わせることを行っている。Exp.2 に重み付の計算式を示す。

$$w_i = \frac{1}{\frac{1}{N}\sum_{n=1}^{N}\left(a_{ni} - \overline{a_{ni}}\right)^2}$$

wi：比較パラメータ i の重み
i：比較パラメータ種別（１～28)
N：単語別サンプル数
n：サンプル番号
a：サンプル別各比較パラメータ
a_n ：各比較パラメータ平均値
Exp. 2　Weight according to comparison parameter

5.3　単語認識システム概要

単語認識システムは全部の単語の一致評価を計算して一致評価の低い順番に並び変えて最も正解に近い動作をしている単語を推測(単語認識)するシステムである。一致評価の計算は、単語辞書の比較パラメータと今撮影したデータの比較パラメータの偏差に２乗対して各パラメータごとの重みを考慮し、さらに各単語のサンプル数で割ったものを全て足すことで計算される。一致評価の値が小さい単語ほど今撮影した動作との動きが近似していると考えられるので、最も評価の低い単語が正解であると判断できる。Exp. 3 に一致評価 D の算出式を示す。

$$D_k = \sum_{i=1}^{28}\left(\left(a_{ki} - b_i\right)^2 \times w_{ki}\right)$$

k：単語種別（単語数はデータベース内を検索して計算）
i：比較パラメータ種別(1~28)
a：単語辞書の比較パラメータ
b：今撮影した映像の比較パラメータ
w：重み
Exp. 3　Agreement evaluation

6.　実験

6.1　単語辞書作成

単語辞書を構成するサンプルデータの作成を行った。
１、単語選定
辞書にするための単語は、Table 2 の２５単語を選んだ。

Table. 2　Selected sign language word

難しい	川	どちら	始まる	長い
飲む	色々	～らしい	風	全部
辛い	来る	大丈夫	松	富士山
おいしい	行く	覚える	ほとんど	東京
甘い	天気	深い	大きい	北海道

２、手話操者の選定と手話単語のレクチャー
手話操者は研究室の学部生、院生、教授等の中から 10 人に協力してもらった。各人とも手話の心得はないため、手話辞典の本と DVD を見せてレクチャーした。

３、ビデオ撮影
手話操者に手話動作をお願いしてビデオカメラで撮影を行った。撮影は制御工学実験室内で行い、１単語につき５人、１人５回づつ(１単語２５動画)撮影した。

４、作成結果
作成したサンプル数は全部で 337 サンプルとなった（Table ３参照)。両腕座標が実際の座標に追従しなかった動画もあるため、サンプル数と撮影数が一致しなかった。

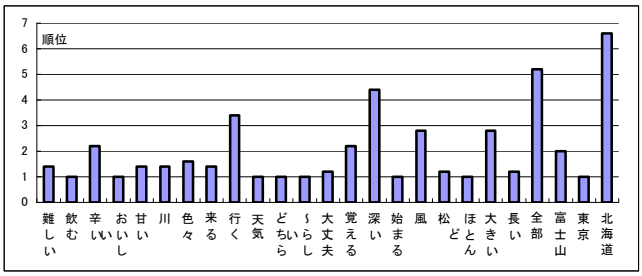
Table. 3　Sample word num

手話単語	フレーム	サンプル	大丈夫	15	16
長い	10	16	東京	23	19
大きい	16	11	天気	14	14
始まる	14	14	おいしい	12	16
ほとんど	16	13	～らしい	18	22
富士山	15	12	飲む	26	24
北海道	21	12	行く	11	10
全部	17	25	甘い	31	17
どちら	32	14	辛い	23	11
深い	17	6	色々	20	9
風	15	5	難しい	23	9
川	11	14	覚える	16	9
来る	8	14	松	17	5

6.2　ビデオ撮影したデータと比較した場合の認識実験

単語辞書データベースが構築した２５単語についてそれ

ぞれ比較プログラムで単語認識を行った。１単語あたりランダムに５つの動画選び、評価の順位を毎回算出して平均の順位を計算した。実験の結果、各単語の平均一致評価は Fig.12 のようになった。認識順位値は低いほど高評価であり、認識順位１位は単語認識に成功したことを示している。



平均順位	2.016位	認識率１位	67.33%
		認識率100%	25単語中8単語

Fig. 12　Word recognition result(video)

各単語の認識率は単語によってバラツキが見て取れるが、25 単語中 17 単語は 2 位以下の平均順位となっているので概ね認識は成功していると言える。平均認識順位が悪い単語の原因をそれぞれ探ると以下のような特徴があった。

① 他に似ている動作の単語が高評価になっている
② 同じ単語でも人によって動作のバラツキが大きい
③ 逆にバラツキが少なすぎて重みが大きくなりすぎる

6.3　リアルタイム撮影データと比較した場合の認識実験

次に、USB カメラで撮影しつつアルタイムに単語認識を行った。撮影は制御工学実験室で行い、３人の手話操者に２５単語の手話をそれぞれ１回ずつ行った。

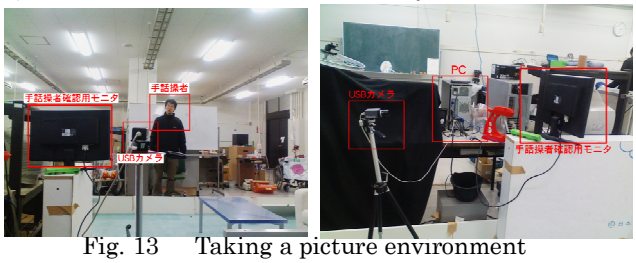
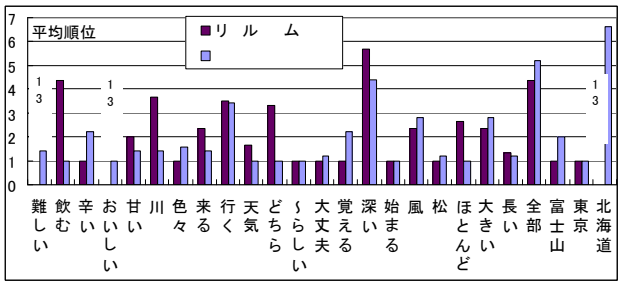


Fig. 13　Taking a picture environment

Fig.13 のような配置でカメラとモニタを設置した。手話操者確認用モニタにはカメラに写る様子がそのまま表示されるため、手話操者はこのモニタを見ることでカメラ内の自分の立ち位置と手が写る領域を確認することができる。Fig.14 にリアルタイム認識を行った場合の単語別平均順位を示す。



手話単語	1	2		大丈夫	1	1	1
難しい				覚える	1	1	1
飲む	7	5	1	深い	6	8	3
辛い	1	1	1	始まる	1	1	1
おいしい				風	3	3	1
甘い	1	3	2	松	1	1	1
川	4	4	3	ほとんど	3	1	4
色々	1	1	1	大きい	3	3	1
来る	1	1	5	長い	2	1	1
行く	3		4	全部	7	5	1
天気	1	1	3	富士山	1	1	1
どちら	3	5	2	東京	1	1	1
～らしい	1	1	1	北海道			

平均順位	3.502	認識率	52.00%
		認識率100%	9　25単語

Fig. 14　Word recognition result(realtime)

実験結果は動画ファイルと比較した場合よりも順位が下がる結果となった。特に「難しい」「おいしい」「北海道」については３人とも 13 位以下の結果となったため認識ができているとは言えなかった。その他の単語については動画ファイル比較の場合と似たような結果となった。

また、３人目の実験結果が他の二人と比べて良好なのは、前の二人の動作を元に動作解析をしてから単語辞書の動きに近づけるようにレクチャーを行ったためである。このことから単語辞書に合わせてレクチャーを行えば高い評価を得ることも可能であると分かった。

7.　結論

本研究では特定の撮影背景や服装に依存せず手話操者にセンサー等を装着しない手話単語認識システムを開発し、その有効性、精度を確認するための実験を行った。その結果、動画ファイルと比較する実験では平均認識順位が２位台だったので、このシステムは有効であると考えられる。しかし、リアルタイム認識だと一部認識できない単語があるため、重み付の計算式などを改良する必要がある。

今後の展望としては、手の形は違うが動作が全く同じ単語にも対応するために、別の手法による補正が必要だと思われる。

参考文献

動画像に対する動き解析/推定
福井大学　情報・メディア工学科　吉田 俊之
http://visix1.fukui-u.ac.jp/researches/motion/index.html
西平、石澤、鳥毛、「手話の自動認識のための基礎的検討－言語認知過程をモデルとした手話動作の認識モデル－」、第１３回ヒューマン・インタフェース・シンポジウム論文集、p.217-220, 1997.
西平、石澤、鳥毛、「手話の自動認識のための基礎的検討－言語認知過程をモデルとした手話動作の認識実験－」、第１３回ヒューマン・インタフェース・シンポジウム論文集、p.221-224, 1997.