手話動画像の意味的特徴点による手話認識技法の考察

宮尾 淳一† 藤原 真里‡
広島大学 総合科学部

１．まえがき

最近の携帯電話においては、カメラや高速な CPU の導入などにより機能および性能において飛躍的な進歩をし、総合的な PDA となりつつある。このような携帯機器の能力を福祉機器に利用することが考えられる。例えば、カメラ付き携帯電話により手話認識が可能になれば、聴覚障害者と健常者のコミュニケーションが容易に行えるようになる。筆者らは、このような状況を最終目標として、ビデオ画像による手話認識の基礎的な検討を行っている。本報告では、基本的な処理の流れを示し、認識処理を高速に行うためのパターンマッチング技法を考察する。

２．手話認識の環境

これまでに手話認識に関して様々な研究が行われており[1][2]、データグローブなどを用いればかなり高い認識率が実現され、かつ、高速な CPU を利用すればリアルタイムに近い認識時間も実現できると予想される。

しかし、上述の携帯電話のようなモバイル環境を想定すると、手話表現者の特別な装備を仮定することができず、特定色のカラー手袋などの利用も避けるべきだと考えられる。さらに、認識を行う機器においても、画像を取り込むカメラは 1 台、かつ、CPU の処理速度もかなり制限されることになる。そのため、複数台のカメラによる立体視も不可能であり、複雑な認識処理も困難になる。本稿では、今後の携帯機器の進展も見込んで次のような環境を想定する。

- 手話表現者は特別の装備を持たず、手指を含んだ上半身を動画像として撮影する。
- 動画は 15fps とし、各フレームは 640×480 ピクセルのフルカラーとする。
- 認識する手話単語は日常生活でよく利用されるものに限定し、すべてを対象にしない。

本稿で提案する手話認識技法は、前述のような入力画像に対して、時間領域・空間領域における処理の削減、パターン解析とマッチング処理の容易化などにより、非力なハードウェアによる認識処理実現を試みる。

A method for Japanese sign language recognition based on the semantic characteristic points in sign language videos
† Jun'ichi Miyao, Faculty of Integrated Arts and Sciences, Hiroshima University
‡ Mari Fujiwara, Faculty of Integrated Arts and Sciences, Hiroshima University

３．時間領域の処理削減

手話は、調動と呼ばれる手指の形と動作により表現されている。そのため、動画像から動作を抽出する必要があり、一つの単語を認識するために、多くのフレームを処理しなければならない。たとえ 15fps としてもその処理は非常に重いものになる。

筆者らは、手話動画像の圧縮に関して既に考察を行っており、単語認識に必要十分な意味的特徴点を明らかにし、これに対応する動画フレーム数は各単語１～３程度であることが分かっている[3]。さらに、意味的特徴点に対応するフレームを近似的に求める比較的単純なアルゴリズムも示している。例えば“好き”という単語に対してアルゴリズムが抽出した２つのフレームを図１に示す(この例では意味的特徴点と一致)。これを用いることにより、処理を行わなければならないフレーム数を１/５程度に削減することができる。

(a) (b)
図１ “好き”の意味的特徴点

４．空間領域の処理削減

認識処理を行うためには、各フレームにおいて、顔と手指の位置と手指の形状が必要になる。認識の精度を高めるためには表情の解析も必要になると考えられるが、処理が増大するので、現時点では表情の処理は行わない。以下では、フレーム内の顔と手指の領域を効率よく求め、必要最小限の領域にすることを考える。処理の高速化のためには皮膚の色情報を利用すること

にする。この方法では、背景などに皮膚と同様な色情報を持つ領域が存在することがあるが、面積や形状が顔や手指と大きく異なり独立しているものに関しては取り除くことができ、顔と手指が重なったり、背景の一部と区別できない場合には輪郭線抽出により分離する。

図１の意味的特徴点フレームから顔と手指の領域を抽出した結果を図２に示す。

(a)　　(b)

図２ 顔、手指領域の抽出

５．認識処理

認識を高精度に行うためには、意味的特徴点における顔と手指の相対的な位置と手指の形状が必要になる。しかし、２次元画像から手指の３次元形状を求めることは、容易ではない。そのため、２次元のままでパターンマッチングを行うことにする。ただし、通常の方法では、大きさ、位置、角度などの正規化が必要となり、処理負荷の増大が懸念される。

そこで、次のような方法により、指の本数や位置の数値的なマッチングとし、高速な認識を目指すことにする。ただし、この方法では１つの画像から得られる情報が十分でない場合もあるが、連続した意味的特徴点すべてにマッチする単語を検索することにより、認識を行う。

(Step1) 顔と手指領域をブロック化し、それぞれの重心を求める。

(Step2) 顔と手指の相対的な位置を求める。

(Step3) 手指領域の水平方向と垂直方向の大きさの 80%の部分を囲む長方形の外周を探索し、立てている指の位置、本数、角度などを求める。

この処理を図２の各画像について行うと、表１に示すような結果が得られる。表 1 より、図２(a)では指が顎付近にあり、指が 2 本上方に立っていることが分かり、(b)では指が顎の下にあり、指が立っていないことが分かる。

表２にこのパターンマッチングに利用するパターンデータの一部を示す。これを用いると表１の例では、２つの解析結果から“好き”という単語であると判定される。

表１ 図２画像の解析結果

	指の本数				掌の向き	顔に対する手指の相対位置
	上辺	左辺	右辺	下辺		
(a)	2	0	1	0	正面	(11, 76)
(b)	0	0	0	0	正面	(45, 108)

表２ 手話認識用パターンデータ（一部）

	動き	指の本数				掌の向き	顔に対する手指の相対位置
		上辺	左辺	右辺	下辺		
好き	下向	2	0	1	0	正面	−20<x<20 0<y<80
		0	0	0	0	正面	−10<x<50 90<y<200
嫌い	下向	0	0	0	0	正面	−20<x<20 0<y<80
		0,1,2	0	1	0	正面	−10<x<50 90<y<200
電話	−	1	0	1	0,1	正面	0<x<150 −20<y<20
悪い	なし	1	0	0	0	正面	−20<x<20 0<y<50
		0	0	1	0	正面	−50<x<20 0<y<50

６．あとがき

本稿では、携帯電話のようなモバイル機器での手話認識を想定して、処理負荷を抑えた手話認識技法を考察した。この方法では、１つの画像を詳細に解析するのではなく、複数の画像から比較的単純な処理で得られる解析結果を総合的に判断して認識することにした。

現在、提案手法の基本的なプログラムを作成しているが、例外的処理やパターンデータなど未完成の部分があるため、これらの実装を行っている。今後、処理時間などの評価を行うと共に、各処理を統合して冗長な処理を除く予定である。また、提案した手法では、認識精度や認識できる単語数などで制限があると思われるが、これに対しては、認識後の文法的解釈などを取り入れることなども検討中である。

なお、本研究の一部は文部科学省科学研究費補助金基盤研究(C)(2)(14550360)で行った。

文献

[1] 澤田,他:”運動特徴と形状特徴に基づいたジェスチャー認識と手話認識への応用”, 情処論, 39, 5, pp.1325-1333, 1998.

[2] 谷端,他:”手話認識のための手指抽出と単語認識”, 信学技法, WIT2001-22, 2001.

[3] 宮尾淳一:”手話における意味的特徴点と手話動画像圧縮への応用”, 信学論 D-I, J84-D-I, 11, pp.1577-1580, 2001.