社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
WIT2005-21(2005-5)

手話のHMM作成のための状態分割

川東 香菜† 白井 良明†† 島田 伸敬†† 三浦 純†

† 大阪大学大学院工学研究科 機械工学専攻
〒 565-0871 大阪府吹田市山田丘 2-1
†† 立命館大学情報理工学部 知能情報学科
〒 525-8577 滋賀県草津市野路東 1-1-1
E-mail: †{kawahigashi,jun}@cv.mech.eng.osaka-u.ac.jp, ††{shirai,shimada}@ci.ritsumei.ac.jp

あらまし 手話の動画像から，手の位置や形に関する特徴を抽出してHMMを作成し，手話単語の認識を行う．顔と手が重なった場合には，隠蔽を考慮して手領域を決定する．得られた手領域より適切な手の特徴を選んで抽出を行う．抽出された特徴を用いて学習を行う際には，HMMの状態数を決定するために画像系列を自動で分割する方法について述べる．作成したHMMを用いた認識実験の結果を示す．

キーワード 手話，HMM，状態，特徴抽出

Segmentation of Sign Language for making HMM

Kana KAWAHIGASHI†, Yoshiaki SHIRAI††, Nobutaka SHIMADA††, and Jun MIURA†

† Department of Mechanical Systems Graduated School of Engineering Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871,JAPAN
†† Department of human and Computer Intelligence Ritsumeikan University
1-1-1 Kusatsu, Shiga 525-8577,Japan
E-mail: †{kawahigashi,jun}@cv.mech.eng.osaka-u.ac.jp, ††{shirai,shimada}@ci.ritsumei.ac.jp

Abstract In this paper, HMM is made by extracting hand features of position and shape, and sign language is recognized. When face and hands are overlapped, hands regions are extracted by considering occlusion. Appropriate features of hands are extracted from hands region. In training phase, image sequences are segmented automatically to decide the number of the states. The experimental result is shown for recognition using HMM.

Key words Sign Language, HMM, State, Feature Extraction

1. はじめに

手話の認識のためには，手の形や動きを認識する必要がある．手の形状データをデータグローブで得るものがあるが[1]，被験者に装着等の負担が生じる．それに比べ，ここで研究を行なっているカメラ画像からデータを得る方法は，被験者に接触する部分がないという利点がある．

画像から手の形状や動きのデータを得る研究は，ジェスチャや手話認識のために広く研究されている．手のシルエットと3次元CGの手とをマッチングすることにより，手の形状を求める方法がある[2]．しかし，背景と手が明確に分離できることを仮定しているので，いろいろな状況に適用するのは難しい．

手話認識のために，肌色の範囲より手の領域を抽出する研究[3],[4]があるが，肌色の範囲は被験者や周りの状況に影響されるので，あらかじめ決められた肌色の範囲では，手の領域を抽出するのは困難である．

一方，手の特徴を抽出して，2次元モデルを用いて認識する方法では，肌色に似た背景の一部により，エラーを含んでいるため，認識が困難である．エラーを含んだ特徴列をロバストに認識する方法として，Hidden Markov Model(以下HMM)がある．HMMは，音声認識，表情認識[5]，ジェスチャー認識[6]，行動認識[7]の分野でよく用いられている．多くの学習サンプルからモデルを構築でき，時間軸の伸縮が可能であるので，スピードの違う手話や，動きが完全に一致しないものでもロバストに認識する可能性がある．

本研究では，画像からの手話単語認識のために，画像系列から適切な手指の特徴を抽出し，その特徴に基づいて画像系列を複数の状態へ分割する方法を中心に述べる．

2. 特徴量抽出のための画像処理

2.1 人物領域の抽出

複雑背景下で撮影された手話画像から人物の領域のみを抽出するには，背景差分を用いる．人物の領域を正確に抽出するためには，人物の影の領域は抽出しないような背景差分を考えなくてはならない．ここでは単純に HSV 色空間で差分をとるのではなく，HSV 色空間を明度の変化をあまり考慮に入れないような色空間に変換してから差分を取る方法 [8] を用いている．

2.2 肌色抽出

人物領域から顔と手の領域を肌色部分として抽出する．しかし，様々な被験者の肌色を判定することは難しい．そこで，撮影初期のフレームから被験者の肌色の情報を取得することによって，様々な被験者に対応する方法を提案する．

被験者には最初に両手を太腿に置いてもらう．このようにしてもらうことで手や顔の位置は想定できているので，この部分から色情報をサンプルすることにより，その被験者の肌色の情報を得ることができる．

肌色の分布は，HS 色空間において正規分布であると仮定して，90% の等確率楕円内に入っており，肌色として取りうる明るさを固定的に与える．その領域にある色を肌色と決定する．前節で説明した人物領域内であり，かつ手の位置に等加速度運動を仮定してその予測位置近傍のみを探索範囲として肌色を抽出する．なお，本研究では顔と手の肌色は微妙な違いがあることを考慮して，それぞれ別の肌色として情報を保存し，そのどちらかで肌色と判定される色を抽出する．

2.3 肘・手首の検出

手話の特徴量を計算する場合に手首の位置が重要である．手首は手領域で肘にもっとも近い点であるとし，肘を発見することで手首を発見する．

肘は円弧状の形状をしていると仮定し，円弧のテンプレートと人物領域の輪郭線とをマッチングすることによって抽出する [8]．円弧の大きさや向きは動きによって変化するので複数のテンプレートを用いて，最もマッチするものを選択する．ただ，輪郭線との単純なマッチングでは，精度良く正確な肘の場所を発見することが難しいので，マッチングには輪郭線画像を距離変換した画像を用いて，円弧テンプレート上の距離の総和が最も小さくなるような場所の円弧の中心を肘の位置とする．そのようにして発見した肘の場所から，手領域（領域の決定方については後述）の中で肘から最も近い点を手首として検出する．

2.4 隠蔽時の判定と処理

手同士が重なった場合や，顔と手が重なった場合には隠蔽の判定が必要である．人物領域内の肌色領域の数と，顔領域の位置はあまり変化しないという仮定のもとでその面積や，前フレームの各領域の位置情報を用いて隠蔽の判定を行う．

また隠蔽が生じていると分かった時には，隠蔽が起こる直前または隠蔽終了直後の肌色領域をテンプレートとして，テンプレートマッチングを行う．

2.4.1 テンプレートの保存

顔と手が重なった時や両手が重なった時には，各領域を正確に抽出するのは難しい．そこで，まず隠蔽が起こる直前の手や顔の形状をテンプレートとして保存しておき，隠蔽中はそのテンプレートを用いてテンプレートマッチングを行う [8]．

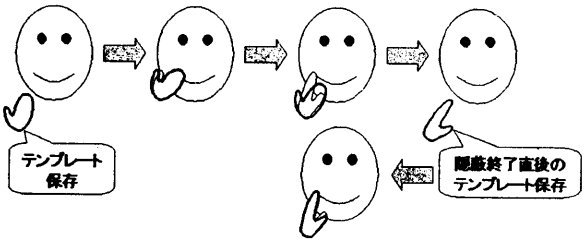


図 1 テンプレートの保存

しかしながら，隠蔽中に手の形が変化することが多く見られることから，隠蔽が終了した直後にも手や顔の形状をテンプレートとして保存する．そして再び，後ろ向きにテンプレートマッチングを行っていき，各フレームで前向きと後ろ向きのテンプレートのどちらがよりマッチしているかを調べる（図 1）．前向きのテンプレートの方がマッチしていると判断された所で後ろ向きのマッチングを終了させ，再び隠蔽終了後のフレームから追跡を行う．図 2 (a) に前向きのテンプレートのみを用いた時の結果を，図 2(b) に前向きと後ろ向きの両方のテンプレートを用いた場合の結果を示す．

2.4.2 手同士の隠蔽

両手を近づけて行う手話では手同士が隠蔽することがある．これを処理する方法について述べる．

手同士の隠蔽の時には保存していた両手のテンプレートを用いて両手の分離を行う．テンプレートマッチングは手の候補となる領域中で，左右の肘から最も近い輪郭上の点を左右の手首位置として，テンプレートの手首位置をその位置に合わせ回転のみのテンプレートマッチングを行い，最もマッチした位置を手の位置とする．

2.4.3 顔と手の隠蔽

顔の前方で手話を行うことは多いので，顔と手の隠蔽はよく起こる．そこで，

- 肌色領域が２つ
- 顔領域のラベルとどちらかの手のラベル番号が一致
- 顔領域の面積が 10% 以上増加

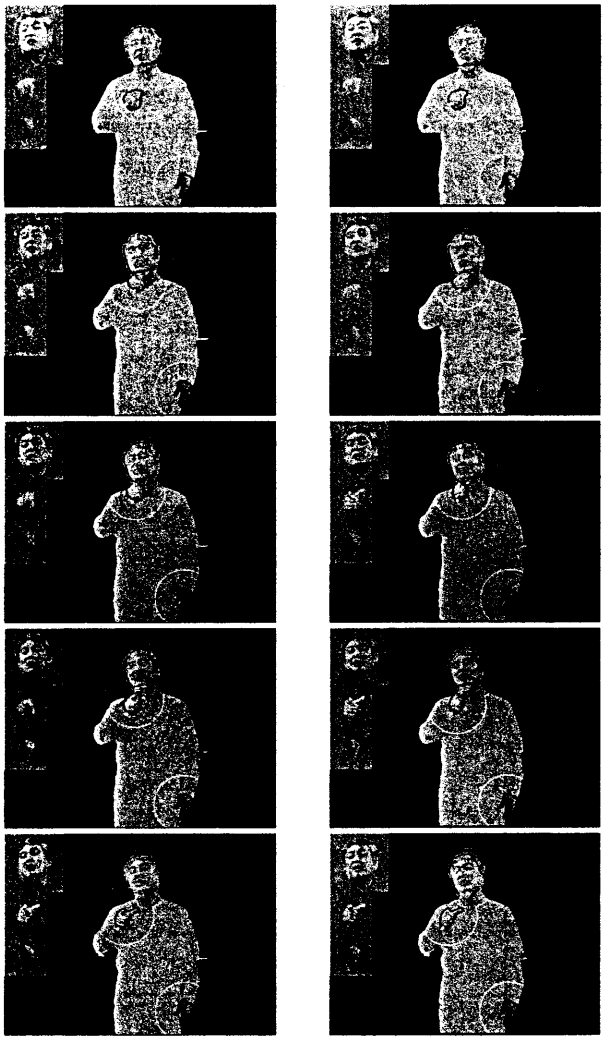
か，

- 肌色領域が２つ
- 前フレームでの手の位置がもう一方の手領域よりも顔領域に近い

という条件を満たしたときには顔と片手の隠蔽が，肌色領域が１つしかないときには顔と両手の隠蔽が起こったと判定する．

この場合にも，手領域と顔領域を分離するためにテンプレートマッチングを行う．隠蔽中，顔テクスチャはほとんど変化しないのでそのテンプレートを現在の顔と仮定する．このことから顔のテンプレートを肌色抽出した画像とマッチングすることによって，現在の顔の位置とすることができる．

手の位置はテンプレートマッチングをすると，顔と手が両方肌色領域であるためにそのままマッチングするだけではうまく



(a) 前向きのテンプレートマッチング　(b) 前向きと後ろ向きのテンプレートマッチング

図 2　テンプレートマッチングの例

抽出できない．そこで隠蔽の起こっている肌色領域と顔のテンプレートにおいて，ブロックごとに相関を計算し，明度相関算 cor の高いブロック (実験的に 0.05 以上) は顔領域であるとすることによってブロックサイズの大きさ程度の精度（本研究ではブロックサイズは 5×5）で手領域を抽出できる（図 3）．相関 cor はあるブロックについて，顔テンプレート画像の明るさ $ftemp(i,j)$, 分離したい画像の明るさ $org(i,j)$ とすると以下の式によって計算している．

$$cor = \sum_{i=0}^{5}\sum_{j=0}^{5}(ftemp(i,j) - org(i,j))^2 \qquad (1)$$

このようにして顔領域を除去した肌色領域内において，保存していた手のテンプレートとマッチングを行う．マッチングはまず顔領域を除いた肌色領域の肘から最も近い点を手首候補点としてそこを中心に，テンプレートを回転させてマッチングを行い，最もマッチした位置を手の位置とする．

3. 手話の特徴量

前章の画像処理によって，両手と顔の領域の位置や形状を得

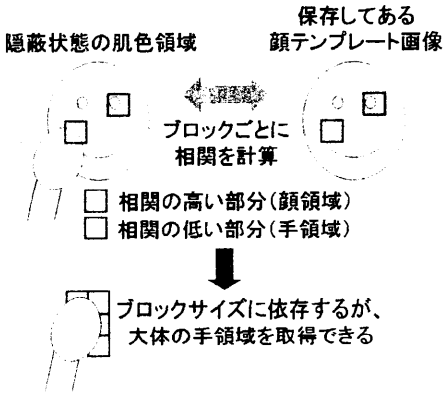


図 3　顔と手の隠蔽時の処理

ることができる．これを HMM によって学習・認識するために，各フレームごとの領域の位置や形状の情報を数値データとして表す．手話の特徴として，「手が動いているときは手の位置の変化が重要であり，手が止っている時には手の形状が重要である.」という傾向がある．この傾向をうまく表現できるような特徴量を定義する．

3.1 手の位置に関する特徴量

手話単語は手の動きや位置によってどの単語かをおおまかに分類できることが多い．そこで，以前は手の位置を表す特徴量として，顔からの距離，顔からの方向，動きの速度，動きの方向を用いていた [8].

手が顔から離れるときには，手の位置はあまり重要ではないことを考慮して，手の位置を極座標系で表す．顔からの距離が一定以上遠いときには変化を少なくするように変更する．顔から手の距離を r とすると特徴量としての顔からの距離 R は，

$$R = \begin{cases} r & (r <= R_a) \\ \sqrt{R_a}\sqrt{r} & (r > R_a) \end{cases} \qquad (2)$$

とする．R_a は顔から離れているとする距離で，本研究では実験的に 150 ピクセルとしている．

動きの方向は，連続的に求めるために前フレームの方向と比較して，角度の差が π よりも小さくなる様に与えていた．しかし，同一単語であっても手の速度が遅い所では，各シーケンスごとに手の軌跡が異なる場合がある（図 4）．それによって図 5 に示すように動きの方向が異なる箇所の影響から，シーケンスごとに角度が約 2π ずれてしまうという問題があった．そこで動きの速度と方向の代わりに，動きベクトルの x, y 方向成分を特徴量として与えることによって図 6 の様に改善することが出来た．

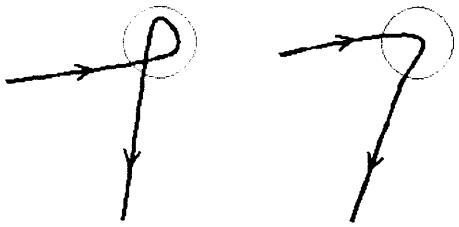


図 4　同一単語で手の軌跡が異なる場合

3.2 手の形状に関する特徴量

手の形状を表す特徴量として，以前は手領域の面積，手領域を楕円近似した時の円形度（長軸/短軸），慣性主軸の方向，手領域の突起数を用いていた [8].

突起数は手首線（手首と重心を結んだ線の垂線）から輪郭線までの距離を測定し，その極値の数とする．このとき，極値とその両隣の谷までの距離を見て，長いものは握りこぶしであるとして突起数 0 とする．

慣性主軸の方向についても動きの方向と同様に角度に約 2π のずれが生じたため（図 7），円形度 $r(0 < r < 1)$ と，慣性主軸方向の x, y 方向成分 $(u, v) : (u^2 + v^2 = 1)$ を用いて $\{u(1-r), v(1-r)\}$ という特徴量を定義する．円形度と慣性主軸方向の代わりに，これらの特徴量を用いることによって，円形度の値が大きい時には，慣性主軸方向の影響を少なくすることが出来る．これによって図 8 のように改善することが出来た．

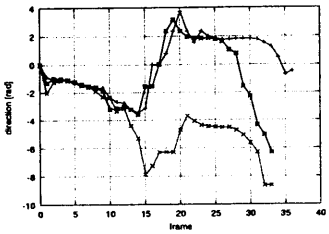


図 5 動きの方向

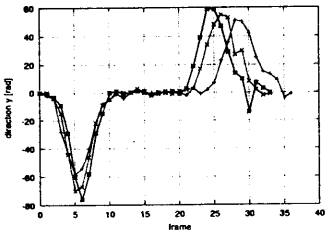


図 6 動きの方向の y 方向成分

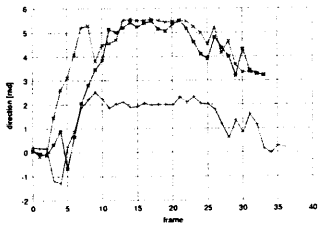


図 7 慣性主軸方向

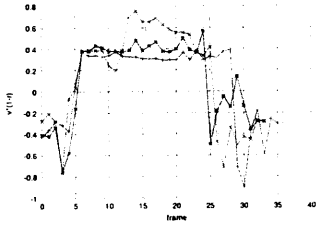


図 8 *v(1-r)*

4. 手話単語の学習と状態数の決定

本章では前章までで抽出した手話特徴量を用いて手話単語を学習する方法について述べる．本研究では学習に用いるアルゴリズムとして，隠れマルコフモデル（Hidden Markov Model(HMM)）を用いる [8]. 本研究で用いる HMM は図 9 に示すような left-to-right モデルである．

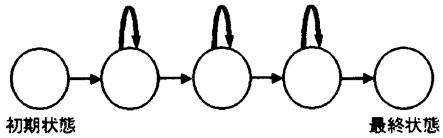


図 9 *left-to-right HMM*

ここでは各単語を学習する時の状態数の決め方を述べる．HMM の出力ベクトルは各状態正規分布に基づく出力確率があるので，よく似た特徴量をもつフレームが同じ状態に属するように，分離したい．そこで手の位置に関する特徴量を用いて画像系列を状態に分割する．

まず片手の場合だけを考える．手が閾値（10pixels/frame）以上の速度で動いているときと，閾値以下の速度になった時をそれぞれ運動区間と静止区間として分割する．速度が閾値以上の区間であっても，速度の極大値が小さく，区間の幅が狭い場合には静止区間とする．ゆっくりと手話を行う人の場合には一旦，静止区間が出来るが，手話のスピードが速い人の場合には静止区間がほとんど出来ずに運動区間に移ることがある．よって手話の始まりと終わり以外での静止区間が，短い場合（3 フレーム以下）に左右の運動区間と併合する（図 10）．また，速度が運動区間内において極小値を取り，その値と両隣の最大値との差が大きい場合（20pixels/frame 以上）には，運動区間をその極小値の点で分割する（図 11）．

手が運動している最中ではあまり方向変化が起こらないが，運動中に速度が一旦，遅くなった時には動きの方向が変化する

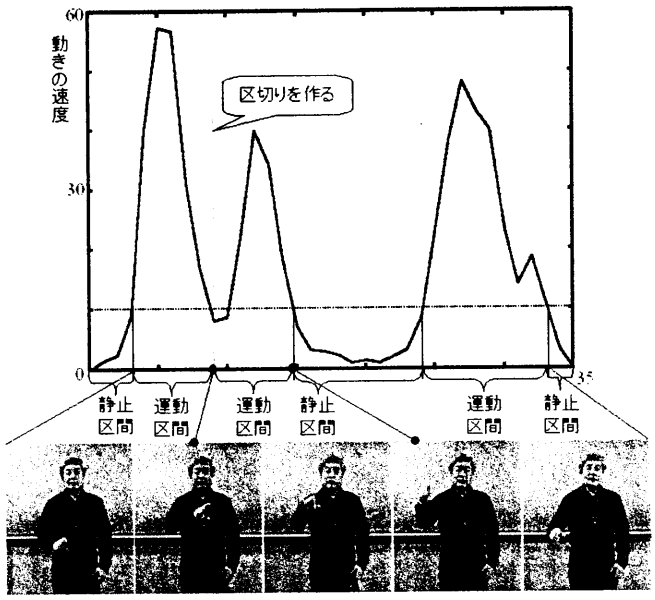


図 10 静止状態を作らずに区切りだけを作る例

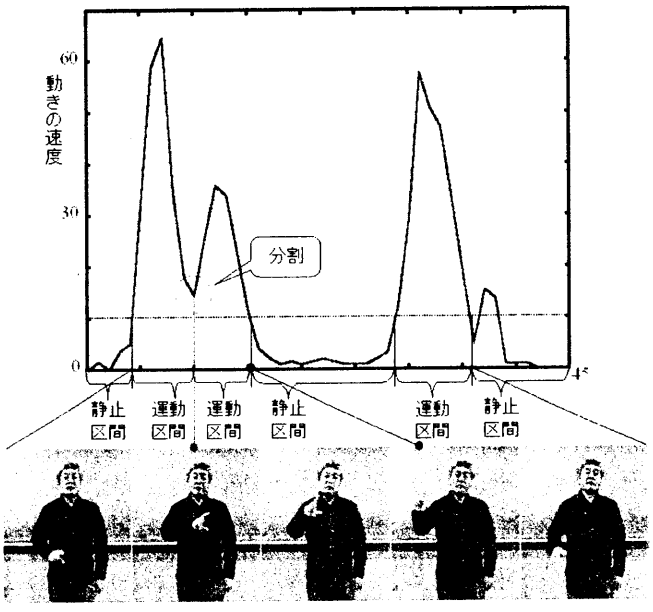


図 11 運動区間内で区切りを作る例

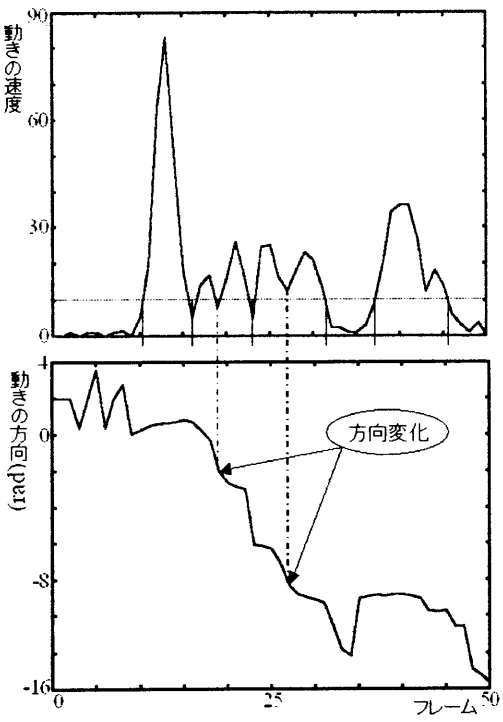


図 12　速度と方向による状態分割の例

ことが多い．そこで手の速度が一定値（5pixels/frame）以上の極小値を取る時に方向が大きく変化（1 フレーム前と 1 フレーム後の方向の差分が 2rad 以上）している場合には，そこを区間分割の候補とする．そして，速度による分割の区切りと前後 1 フレーム以上離れているフレームで，方向変化が起こっている場合には，そこで区間を分割する．

このようにして分割したそれぞれの区間を一つの状態とする．状態分割の一例を図 12 に示す．

両手で行う手話については上記の方法によって，それぞれの手について状態への分割を行い，両方の変化を合わせて状態遷移とする．右手と同時または近いフレームで左手の状態遷移が起こっている場合には，右手と同時に状態遷移が起こったとする．

同じ単語であっても，手話を行う度に状態数が異なる場合がある．図 13(a) では，手話を行う際に静止区間が 4 つ出来るが，図 13(b) では静止区間が 3 つしか出来ない．静止区間が出来るか否かは，同じ被験者でもその度ごとに変化し，また個人の手話を行うスピードによっても変わってくると考えられる．したがって手話を行う度に静止区間の数が異なる単語の場合には，それぞれ静止区間の数が違うモデルを作り，同じ単語として登録する．

他に動きの方向変化が各シーケンスごとに異なる場合や，テンプレートマッチングの際に手の位置がずれる場合に，状態数が揃わない例が存在した．今後，更に隠蔽時の処理の改善を行い，様々な状況に応じたモデルを作成することが課題となる．

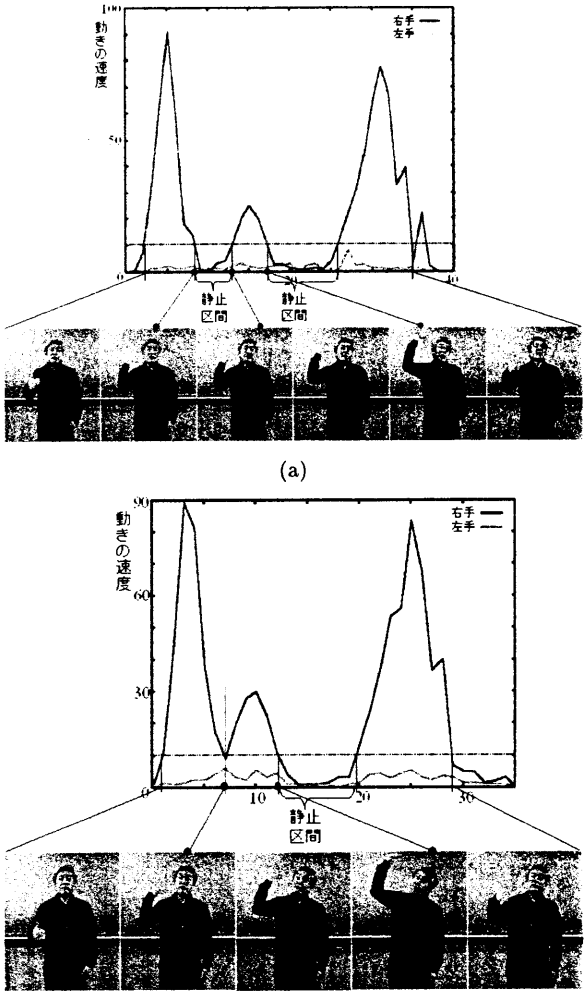


図 13　静止状態の数が異なる場合

5. 実　　験

5.1　実験に用いた手話画像

本研究では，手話熟練者による手話画像を用いた．撮影は大学内の講義室で行い，背景に特に暗幕などを用いず自然な背景とした．なお特徴量抽出（画像処理）は現在 1 フレームあたり数秒かかるため，撮影した画像（30 フレーム/秒）を 1 フレームおきに処理する．実験に用いたサンプルは 45 単語であり，1 単語につき 3 シーケンスある．

5.2　手話単語の認識結果

4. で述べた方法によって状態分割を行ったところ，3 シーケンスとも状態数が揃ったものは 21 単語（両手 8 単語，片手 13 単語）であった．また，静止区間の数が異なることから 2 シーケンスでは状態数が揃うが，残り 1 シーケンスでは状態数が異なるものは 5 単語（両手 2 単語，片手 3 単語）であった．この 26 単語のみを用いて，両手と片手の手話で別々に認識を行った．それぞれの単語について組合わせを変えながら 3 シーケンスのうち 2 つを学習用に，残りの 1 つを認識用に用いた．状態数の異なる単語ではモデルを 2 種類作成した．認識結果を表 1，2，3 に示す．表 2，3 で，認識に成功したものは○で，失敗し

たものは誤認識した単語を示している．

表 1 に示す様に両手，片手の手話共に高い認識率を得ることが出来た．誤認識した例として「厚い」と「赤」の顔からの距離と $v(1-r)$ の，各シーケンスとの比較を図 14 と図 15 に示す．認識データを赤で，学習データを緑と青で表示している．また，それぞれの画像シーケンスを図 16 と図 17 に示す．図から分かるように各シーケンスにおいてデータ間にばらつきが生じ，間違った単語が認識結果として上がってしまったと考えられる．

今回はサンプル数が少なかったため，今後はより正確に状態分割を行い，学習に用いることの出来る単語数を増やす必要がある．また，被験者を増やしてさらに学習を行うことによって，ばらつきの影響を減らすことも重要である．

表 1: 認識結果

	両手	片手
1 回目	9/10	13/16
2 回目	10/10	16/16
3 回目	10/10	16/16
合計	29/30	45/48
認識率	0.97	0.94

表 2: 両手の手話

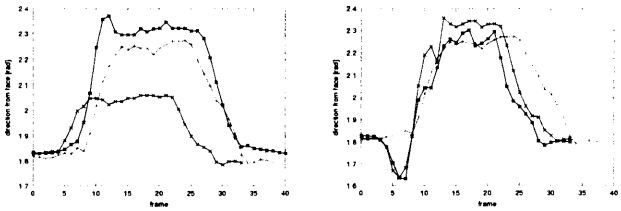
	1 回目	2 回目	3 回目
セーター	○	○	○
めがね	○	○	○
皮	○	○	○
薄い	○	○	○
長い	○	○	○
夏物	○	○	○
暖かい	○	○	○
寒い	○	○	○
短い	○	○	○
値上げ	寒い	○	○

表 3: 片手の手話

	1 回目	2 回目	3 回目
かばん	○	○	○
赤	○	○	○
黒	○	○	○
白	○	○	○
絹	○	○	○
厚い	赤	○	○
円	○	○	○
胸	○	○	○
頭	○	○	○
安い	○	○	○
高い	○	○	○
～はどこですか	○	○	○
～していいですか	○	○	○
ネクタイ	白	○	○
背が高い	○	○	○
背が低い	背が高い	○	○

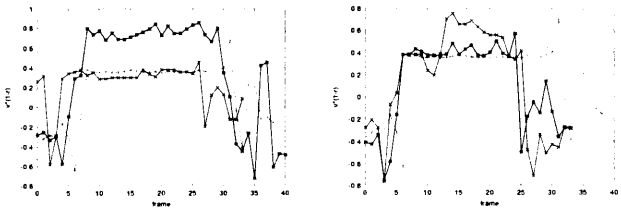
6. おわりに

本研究では画像から隠蔽を考慮した顔や手の領域を抽出し，手の動きの速度と方向によって画像系列を状態に分割する手法を提案した．また，その分割結果による状態数を用いてモデルを作成し，認識実験を行った結果，高い認識率が得られた．今後は手の形状に関する特徴も用いて状態分割を行う必要がある．また手の形状の特徴は画像処理のエラーを含みやすいことから，テンプレートマッチングによる手法の更なる改善が課題



(a)「厚い」同士の比較　(b)「厚い」と「赤」の比較

図 14　顔からの距離の比較



(a)「厚い」同士の比較　(b)「厚い」と「赤」の比較

図 15　$v(1-r)$ の比較

図 16「厚い」

図 17「赤」

となる．今後，さらにサンプル数や被験者を増やすことによって認識率の向上を目指す．

文　献

[1] H.Sagawa, M.Takeuchi, "A Method for Recognizing a Sequence of Sign language Words Represented in Japanese Sign Language Sentence", *Proc. Int. Conf. Automatic Face and Gesture Recognition (FG2000)*, pp. 434–439, 2000.

[2] N. Shimada, K. Kimura and Y. Shirai, "Real-time 3-D Hand Posture Estimation based on 2-D Appearance Retrieval Using Monocular Camera", *Proc. Int. WS. on RATFG-RTS (satellite WS of ICCV2001)*, pp. 23–30, 2001.

[3] K.Imagawa, "Color-Based Hands Tracking System for Sign Language Recognition", *FG1998*, pp. 462–467, 1998.

[4] K.Imagawa, H.Matsuo, R.Taniguch, and D.Arita, "Recognition of Local Features for Camera-based Sign Language Recognition System", *FG2000*, pp. 849–0853, 2000.

[5] 坂口，大谷，岸野："隠れマルコフモデルによる顔画像からの表情認識"，テレビジョン学会誌，Vol. 49, No. 8 1995.

[6] Takio Kurita, Satoru Hayamizu, "Gesture Recognition using HLAC Features of PARCOR Images and HMM based Recognizer", *FG1998*,pp. 422–427,1998.

[7] 大和，大谷，石井："隠れマルコフモデルを用いた動画像からの人物の行動認識"，電子情報通信学会論文誌,Vol. J76-D-II, No. 12 1993.

[8] 金山，白井，鳥用："HMM を用いた手話単語の認識"，信学技報，Vol. 104, No. 93, pp. 21–28, 2004.