K-3

主成分分析を用いたHMMによる手話認識

Sign Language Recognition Based on HMM Using Principal Component Analysis

柳生 雄午† 宮島 千代美† 徳田 恵一† 北村 正†
Yugo YAGYU Chiyomi MIYAJIMA Keiichi TOKUDA Tadashi KITAMURA

1. まえがき

近年, 聴覚障害者の対話支援を目的とした手話認識システムに関する研究が行われるようになってきた. 手話認識における手話のモデル化には, 時系列パターンの表現に適した隠れマルコフモデル (HMM) が多く用いられている. また, 手話の特徴抽出に関しては, データグローブのような機器を用いて特徴点を直接入力したり, 各画像フレームから手の位置や大きさなどを抽出して特徴量とするモデルベース法に基づいた手法が多く用いられている [1]–[3].

動画像の認識法には, モデルベース法の他に画像ベース法と呼ばれる手法がある. 画像ベース法は, 画像の画素情報を特徴ベクトルとするもので, 特徴ベクトルのサイズが大きくなってしまうため, 多くの場合は何らかの特徴抽出法によって次元数の削減が行われる. 我々はこれまで, 画像ベースに基づくジェスチャー認識の研究を行ってきたが, 画像の特徴抽出に主成分分析 (PCA) を用いることによって, 大幅な次元数削減となるとともに, 高い認識率が得られることを確認した [4].

本研究では, これを手話認識システムに拡張し, RWC 人間動作理解大規模ジェスチャーデータベース [5] を用いて画像ベースでの認識法について検討する.

2. 主成分分析 (PCA) による特徴抽出

1 枚の画像の全画素値を 1 列に並べた M 次元ベクトルを $\boldsymbol{x} = [x_1, x_2, \cdots, x_M]^t$ とする. N 枚の画像 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ を用意し, 画像 $\boldsymbol{x}$ から平均画像 $\bar{\boldsymbol{x}}$ を引いたものを $\hat{\boldsymbol{x}} = \boldsymbol{x} - \bar{\boldsymbol{x}}$ と表す. このとき, N 枚の画像から得られる行列 $\hat{X} = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_N]$ に対して PCA を行うことによって正規直交基底 $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N]$ が得られる. $\boldsymbol{U}$ は固有ベクトルとも呼ばれ, 再構成画像 $\hat{\boldsymbol{x}}$ は $\boldsymbol{U}$ の線形結合によって, $\hat{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{y}$ のように表すことができる. ここで, 主成分スコア $\boldsymbol{y} = [y_1, y_2, \cdots, y_N]^t$ は原画像 $\boldsymbol{x}$ を表す特徴量と考えることができる. 従って, 再構成に使用する $\boldsymbol{U}$ と $\boldsymbol{y}$ の次元 n を N より小さくすることで, 特徴空間の次元を圧縮することができる.

3. 手話認識実験

3.1 データベース

実験で用いるデータベースには男性 2 名 (m1, m2), 女性 2 名 (f1, f2) の 64 種類の手話文章が 2 回ずつ収録されている. 画像は正面 (a) と被験者の左側 (b) の 2 箇所から撮影されており, 立体的なデータの扱いが可能である. 本実験では, 正面画像 (a) のみを用いる. データベースには表 1 に示す 64 種類の手話文章が収録されており, 計 38 種類の手話単語 [6] が存在する. また, 各文章は 5〜6 個の単語で構成されている. 画像サイズは 320×240 画素, フレームレートは 1/30 秒である.

†名古屋工業大学 知能情報システム学科,
Dept. of Computer Science, Nagoya Institute of Technology

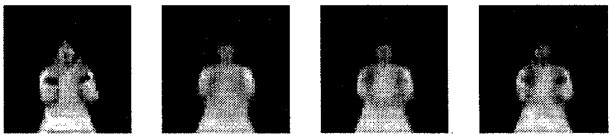
表 1: 文章に含まれる単語

文章番号	収録単語
1〜16	いつも, 毎日, 放課後, 夜, 法律, 文学, 勉強, 宿題, 〜する
17〜32	公園, 銀行, 裏, 近所, 家, アパート, 住む, 暮らす, 新しい
33〜48	関東, 北九州, 北海道, 東北, 晴れ, 曇り, 雨, 雪, しかし, 〜です
49〜64	春, 秋, スペイン, マレーシア, ベトナム, メキシコ, 旅行, 研修, どちらですか, 〜ですか

3.2 実験条件

全 64 文章を単語の出現頻度が同じになるように A セット, B セットの 2 つに分割し, A セットを学習用, B セットをテスト用データとした. また, 4 名のうち男女各 1 名ずつ (計 2 名) を学習データ, 残りの 2 名をテストデータとする Jack-knife 法により, 男女の組み合わせを変えて 4 回認識を繰り返した.

前処理として, 原画像を左右の画素値の和が同程度になる位置を基準に左右に 120 画素ずつ, 240×240 画素に切り出した. 次に, 学習データからランダムに選んだ 1000 枚の画像に対して PCA を行い, 求められた主成分スコア 10, 20, 50 次元を特徴量として用いた. 図 1 に, これらの次元数に関して再構成画像 $\hat{\boldsymbol{x}}$ を求め, 平均ベクトル $\bar{\boldsymbol{x}}$ を加えたものを示す. 再構成画像は全体的に顔や手の輪郭がぼやけた画像となるが, 大まかな動作は理解することが可能である. なお, これまで我々が行ってきたジェスチャー認識の実験では, 主成分スコアの次元 $n = 20$ の場合に最も高い認識率が得られている. 認識実験は表 2 の条件で行った.



原画像 n=10 n=20 n=50

図 1: 原画像と再構成画像

表 2: 実験条件

モデル数	手話単語 38 モデル
状態数	各モデル 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 状態
混合数	各状態 1 混合
特徴量	主成分スコア 10, 20, 50 次元, Δ, Δ^2
実験方法	学習データ： 2 名 × 32 文章 × 2 回 テストデータ : 2 名 × 32 文章 × 2 回

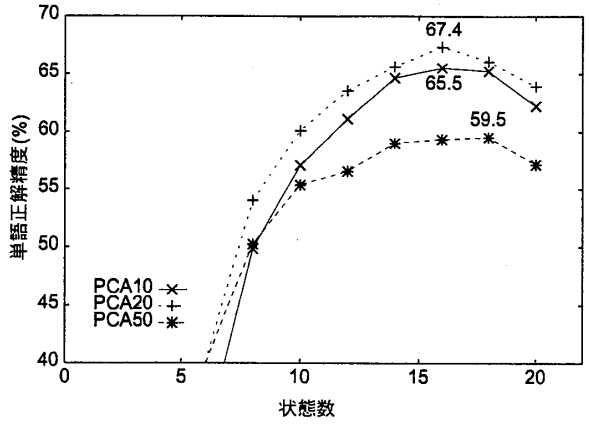


図 2: 状態数と単語正解精度

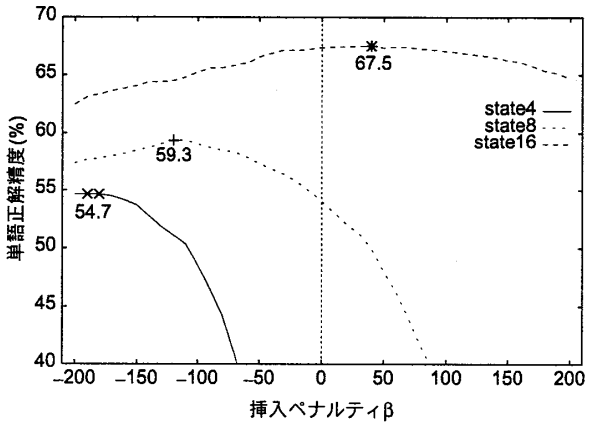


図 3: 挿入ペナルティと単語正解精度 (n=20)

3.3 実験結果

HMM の状態数と単語正解精度の関係を図 2 に示す. ただし, 単語正解精度 Acc(%) は, 式 (1) のように計算される.

$$Acc = \frac{H - D - S - I}{N} \times 100 \qquad (1)$$

ここで, N は認識単語の総数, H は正解単語数, D は削除誤り単語数, S は置換誤り単語数, I は挿入誤り単語数を表している.

ジェスチャー認識の場合と同様に, 全体的に主成分スコアの次元数が 20 のときに良い結果が得られた. また, HMM の状態数が 16 の場合に最も良い結果が得られ, 最高で 67.4% の単語正解精度となった.

一方, 状態数が少ない場合では正解精度が悪くなってしまうという傾向がある. そこで, 挿入ペナルティ β を導入する. $\beta \times$ (文章に含まれるモデル数) を認識時の対数尤度に加えることによって, 尤度を補正する.

実験では, β の値を -200〜200 まで 10 刻みで変化させて, 正解精度の変化を調べた. 図 3 に PCA の次元数 20 での状態数 4, 8, 16 における挿入ペナルティと単語正解精度の変化の様子を示す. また, 各状態ごとに最適な挿入ペナルティを与えた場合の状態数と単語正解精度の変化を図 4 に示す. 図 3 より, 状態数が 4, 8 と少ないところでは, 挿入ペナルティの値を負にすることによって, 大幅な正解精度の向上が見られる. 逆に, 状態数 16 の辺りでは, 挿入ペナルティが 0 の付近で最も正解精度が高くなっているため, 正解精度に大きな改善は見られなかった. 最終的に図 4 に示すように, 挿入ペナルティを用いた場合でも PCA の次元数が 20, 状態数が 16 の場合の正解精度が最も高く, 67.5%となった.

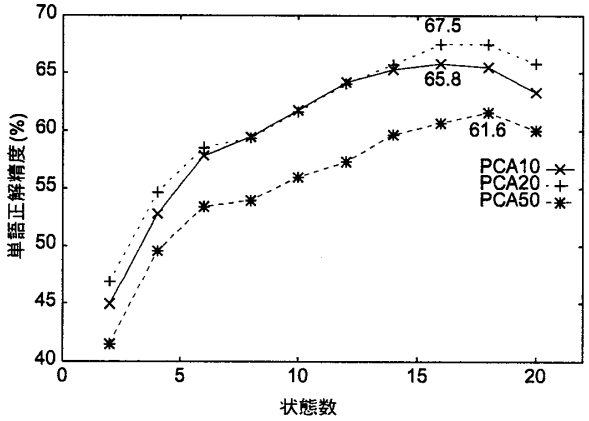


図 4: 状態数と挿入ペナルティを加えた単語正解精度

4. むすび

本研究では, 不特定話者 (学習 2 名, 認識 2 名) による 38 単語の連続手話認識実験を行い, 画像ベースで 67.5%の単語正解精度を得た.

今後の課題としては, 2 方向からの画像を用いた認識, モデルベース法との比較, 手話文法を考慮した認識などが考えられる.

謝辞

本研究の一部は, 立松財団研究助成により行われた.

参考文献

[1] B. Bauer and H. Hienz, "Relevant features for video-based continuous sign language recognition," Proc. FG2000, pp.440–445, Mar. 2000.

[2] T. Starner, J. Weaver and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," Proc. PAMI'98, vol.20, no.12, pp.1371–1375, Dec. 1998.

[3] T. Kobayashi and S. Haruyama, "Partly hidden Markov model and its application to gesture recognition," Proc. ICASSP'97, vol.6, pp.3081–3084, Apr. 1997.

[4] 中谷 博美, 酒向 慎司, 徳田 恵一, 北村 正, "固有ジェスチャーを用いた HMM に基づくジェスチャー認識," 信学総大, D-12-116, p.283, Mar. 2001.

[5] マルチモーダルデータベースサブ WG, "RWC 人間動作理解大規模ジェスチャーデータベース," http://www.rwcp.or.jp

[6] 丸山 浩路, "初めての手話の本," 祥伝社, 1993.