

## インタラクションのための人物計測とジェスチャ認識

岩井 儀雄, 谷内田 正彦, 萬上 圭太, 樹木 義道  
大阪大学大学院基礎工学研究科

現代の計算機は音声や映像などの各種メディアを扱えるようになり格段に進歩した。しかし、人間の意図や感性にマッチした情報機器インターフェイスはまだ実現していない。そこで、身体ジェスチャ、表情や視線変化などの人に優しいマンーマシンインターフェイスの実現を目指して、非接触な観測が可能であるカメラを利用したジェスチャを認識手法を提案する。

## Human Motion Measurement and Gesture Recognition for Interaction

Yoshio Iwai, Masahiko Yachida, Keita Manjoh, Yoshimichi Ueki  
Graduate School of Engineering Science, Osaka University

Computer can directly treat with audio and visual information at present. Human-computer interface, however, is not realized which adapts to human intension or kansei. Sensing of human motion is very important for human-computer interactive applications. A vision system is suitable for human-computer interaction since this involves passive sensing and the system can estimate the motion of the user without any discomfort for the user. In this paper, we propose methods for gesture recognition by using a vision system to actualize such interface.

### 1 はじめに

人間の意図や感性にマッチした情報機器インターフェイスを実現するための基礎的応用的研究を目的として、日本学術振興会未来開拓学術研究推進事業「感性的ヒューマンインタフェイス」(研究推進委員: 原島博東大教授)の研究が始まっている。

我々のグループでは、「インタラクションによる相乗効果を用いた感性創発世界の構築」(プロジェクトリーダー: 谷内田正彦大阪大教授)の元で、インタラクションにより触発され、増幅される感性情報を媒介とする感性創発世界(システム)を構築することを目的としている。具体的には、共存感を持った上での臨場感あふれる対象シーンの提示環境の構築、身体ジェスチャ、表情や視線変化による人に優しいマンーマシンインターフェイスの実現などを目指している。特に、本稿では、身体ジェスチャ認識において現時点におけるプロジェクトの進捗具合を報告する。

身体ジェスチャを情報機器の入力とする場合には、違和感を生じさせないようにできるだけ身体拘束が少ない方が望ましい。そこで、我々は非接触観測が可能であるカメラを利用して身体ジェスチャを計測する手法を開発している。

### 2 カメラによる身体計測関連研究

装着物を使用しない、カメラによる動きの計測[1, 2, 3, 4]はセンサを意識していない対象の計測を可能とし、非接触型であることから適用可能な状況は広範囲に渡ると考えられる。センサにカメラを用いる研究としては、単眼カメラを用いたジェスチャの認識[5]がある。この手法は手を色と位置情報より時系列で追跡することによってジェスチャをパターンに分類している。しかし大掛かりな装置を必要としない利点はあるが、オクルージョンを考慮していないために応用範囲は限られている。

近年ではジェスチャの認識に音声認識の分野で成果を上げている隠れマルコフモデル(HMM)を用いる研究も多く行なわれている。Yamato[9]は離散HMMとベクトル量子化の技術を用いて、テニスの6種類のスウィングの前処理を行った画像シーケンスに対して認識を行った。また、Schuster[10]は、同様に離散HMMを10種類の人の動作認識に利用し、前処理を簡略化することで実時間認識の可能性を示唆した。HMMは動画像だけでなく顔認識やハンドライティング認識といった単画像認識にも用いられている[11, 12]。これらは、HMMを階層的に用いた疑似

2次元 HMM を利用したものである。Müller[13] はこれをさらに拡張した疑似3次元 HMM によって、動きのあるジェスチャー認識だけでなく、ポーズだけのジェスチャー認識も実現した。しかし、これらのうち、動画像における動作認識については実時間化の実現に至っていない。

### 3 動きモデルを用いた人物の姿勢推定

本稿では高速で応用範囲の広いジェスチャー動作の計測方法を提案する。対象者への装着物は使用せず、センサにカメラを用いて頭、手の3次元位置を推定し、人の形状モデルと背景差分によって切り出した人物領域とのモデルマッチングを行なう。また人の動きを関節角度によって時系列で表現した動きモデル[6]を統計的なデータに拡張し、個人差を反映した、より真値に近い姿勢を推定する。

#### 3.1 人の形状モデルと座標系

人の形状モデルと、その座標系を図1に示す。関節の可動範囲には表1のような制限を与える。人の形状モデルは頭、肩、肘、手の特徴点として表現され、特徴点間のリンクの長さは既知とする。

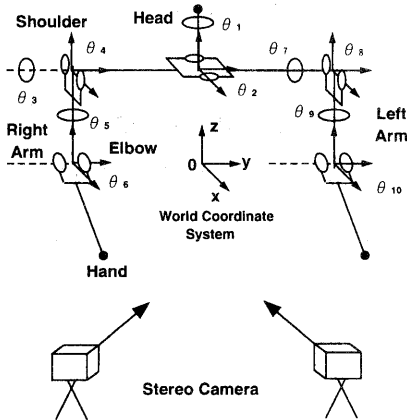


図1: 人の形状モデルと座標系

#### 3.2 統計的動きモデル

本稿では人の姿勢を図1で示した合計10関節の角度で表し、ある時刻における人の状態を式(1)で表現し、この20次元で張られる空間を $\Theta$ 空間とする。また図1より、右手、左手の頭に対する相対位置、速度の状態を表す特徴ベクトルを12次元観測ベクトル空間( $X$ 空間)で式(2)のように表す。

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_{10}, \dot{\theta}_1, \dots, \dot{\theta}_{10}\} \quad (1)$$

表1: 形状モデルの関節可動範囲

| $\theta_i(\text{deg.})$ | min  | max |
|-------------------------|------|-----|
| $\theta_1$              | -180 | 180 |
| $\theta_2$              | -15  | 15  |
| $\theta_{3,7}$          | -45  | 180 |
| $\theta_{4,8}$          | -90  | 90  |
| $\theta_{5,9}$          | -90  | 90  |
| $\theta_{6,10}$         | 0    | 180 |

$$X = \{x_r, y_r, z_r, x_l, y_l, z_l, \dot{x}_r, \dot{y}_r, \dot{z}_r, \dot{x}_l, \dot{y}_l, \dot{z}_l\} \quad (2)$$

$X$ と $\Theta$ の関係式は順運動学問題を解くことにより行列 $A$ を用いて式(3)で表せる。

$$X = A\Theta \quad (3)$$

入力される $X$ 空間上の点は式(3)を逆に解くことによって $\Theta$ 空間へ変換することにより直接姿勢を推定できるが、解は一意には決まらない。そこで本稿では予め統計的な人の動きモデルをもっていき、モデルとの類似度を用いて姿勢を推定する。動きモデルは推定対象となる各ジェスチャーを実行者が複数回実行して作成する。

また作成した全ての動きモデルを式(3)によって $X$ 空間へ変換し、変換前後で空間間での点と点との対応を記録する。すなわち各動きモデルは「ジェスチャー種類」、「 $X$ 空間座標」、「 $\Theta$ 空間座標」を保持している。本稿では表2に示すように「バンザイ」、「パンチ」、「バイバイ」の3つのジェスチャーに関して各3系列ずつ動きモデルを作成した。これらを標準偏差5度のガウス分布に従って分散させ、各々100個、合計43,500個の動きモデルを作成した。

表2: 動きモデルの数

| Gesture | Series.1 | Series.2 | Series.3 | Total   |
|---------|----------|----------|----------|---------|
| BANZAI  | 32       | 28       | 30       | 90×100  |
| PUNCH   | 40       | 33       | 34       | 107×100 |
| BYEBYE  | 82       | 80       | 76       | 238×100 |

#### 3.3 姿勢推定のアイデア

本稿における姿勢推定システムを式(4),5に示す。

$$\hat{\Theta}(T+1) = \begin{bmatrix} I & \Delta \times I \\ 0 & I \end{bmatrix} \hat{\Theta}(T) \quad (4)$$

$$\hat{x}(T) = A\hat{\Theta}(T) + u(T) \quad (5)$$

ただし、 $\hat{\Theta}(T)$  は時刻  $T$  における推定姿勢、 $I$  は  $10 \times 10$  の単位行列、 $\Delta$  は  $1/30(\text{sec})$ 、 $\hat{x}(T)$  は観測値、 $A$  は順運動学行列、 $u$  はノイズを表す。本稿では図 2 のように姿勢を推定する。

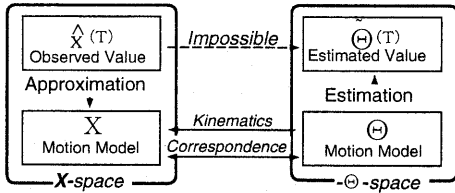


図 2: 姿勢推定の概要図

画像上から頭に対する両手の 3 次元相対位置、速度を観測し、12 次元観測ベクトル空間で入力  $\hat{x}$  を構成して  $\hat{x}$  と  $X$  空間上での動きモデル  $X$  との類似度を調べ、その対応する  $\Theta$  空間上の動きモデル  $\Theta$  から姿勢を推定する。

### 3.4 姿勢推定アルゴリズム

姿勢推定処理の流れを図 3 に示す。ただし、本稿では姿勢のパラメータ最適化については未実装である。

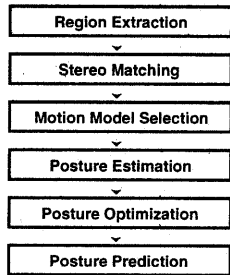


図 3: システムの流れ

#### 3.4.1 顔、手の抽出と 3 次元位置推定

2 台のカメラ画像情報を統合して顔、手の領域を抽出する。視点の異なる 2 つのカメラ画像について色とカメラ座標によって顔、手領域を決定し、ステレオマッチングによって頭、手の 3 次元位置を推定する。

#### 3.4.2 動きモデルとクラスタリング

入力  $\hat{x}$  と動きモデル  $\Theta$  との類似度を計算するために、全て ( $N$  個) の  $\Theta$  を  $r$  個にクラスタリングし、パターンに分類する。実際のクラスタリングの方法に

ついてはまず、全ての  $\Theta$  に関して式 (3) を適用することによって  $X$  を計算する。次に各  $X$  について  $k$ -平均法によって  $r$  個にクラスタリングする。また  $k$ -平均法の評価基準となる距離にはユークリッド距離を用いるものとし、特徴量は予め正規化しておく。また、位置と速度には 5:1 の重みをかけた。

実際にクラスタリングを行なう場合、各クラスタの性質はその総数  $r$  個に大きく依存する。 $r$  が大きければクラスタはコンパクトになるが、冗長なクラスタも多くなる。また  $r$  が小さければ冗長なクラスタは少なくなるが、分散は大きくなる。本稿では表 2 に示した通り、435 個のシーンがあるので、クラスタ総数は最大で 435 個あれば良い。しかしこれらは重複したシーンも含むため、冗長なクラスタが存在する。そこで本稿ではクラスタがどの程度コンパクトにまとまっているかを知るための評価関数として式 (6) を用いて  $r$  を決定する。

$$J(r) = \left( \frac{J_X(r)}{J_\Theta(r)} \right) = \left( \frac{J'_X(r)/J'_X(435)}{J'_\Theta(r)/J'_\Theta(435)} \right) \quad (6)$$

$$J'_X(r) = \sum_{i=1}^r \sum_{X \in \omega_i} (X - m_i)^t (X - m_i) \quad (7)$$

$$J'_\Theta(r) = \sum_{i=1}^r \sum_{\Theta \in \omega_i} (\Theta - M_i)^t (\Theta - M_i) \quad (8)$$

$$m_i = \frac{1}{n_i} \sum_{x \in \omega_i} x \quad (9)$$

$$M_i = \frac{1}{n_i} \sum_{\Theta \in \omega_i} \Theta \quad (10)$$

ただし、 $n_i$  はクラス  $\omega_i$  のサンプル数を表す。

#### 3.4.3 ジェスチャ推定

時刻  $T$  において頭に対する手の 3 次元相対位置が測定されても、自由度が足りないため各関節角度を一意に決めることはできない。そこで観測された手の相対位置と速度から 12 次元観測ベクトル空間上で入力  $\hat{x}$  と各クラスタとの類似度を計算する。本稿では 3.4.2 節で述べたクラスタリングによって  $r$  個のクラスタ  $\omega_i$  ( $0 \leq i \leq r$ ) が生成されたときに、入力  $\hat{x}$  は式 (11) に表わされるユークリッド距離  $D(\hat{x}, m_i)$ :

$$D(\hat{x}, m_i) = (\hat{x} - m_i)^t (\hat{x} - m_i) \quad (11)$$

を最小にするような  $\omega_i$  に最も類似し、このクラスタに所属しているとする。類似度の計算にはマハラノビス距離を利用するのも有効だが、処理時間が多くかかる上に大きな精度の向上は見られなかったため本稿ではユークリッド距離を用いた。

次に、求められた所属クラスを $\omega_{q_0}$ として、残り全てのクラスについて $D(m_{q_0}, m_j)$  (ただし $q_0 \neq j$ ) が小さい順番に $\omega_{q_1}, \omega_{q_2}, \dots, \omega_{q_{r-1}}$ とすると式(12)に表される $q$ を構成できる。

$$q = \{\omega_{q_0}, \omega_{q_1}, \dots, \omega_{q_{r-1}}\} \quad (12)$$

本稿では実際に入力 $\hat{x}$ が得られたとき、各クラスに対する $\hat{x}$ の類似性の順位は $q$ で表されるとする。式(11)を計算すると厳密には異なるが、本稿のようにクラス数が多い場合、この近似は十分に成り立つと考えられる。これによって $\omega_{q_0}$ が求まれば、 $q$ が一意に決まるため、予め $\omega_{q_0}$ に対応する $q$ をオフラインで計算できる利点がある。

入力 $\hat{x}$ が所属するクラス $\omega_{q_0}$ の内部は各ジェスチャの動きモデルデータが混在している。この比率がジェスチャ $G_j$  ( $j$ はジェスチャの種類)の生起確率と考えられるが、本稿ではノイズを考慮して過去 $K$ フレームの生起確率を平滑化して最終的な生起確率を算出する。時刻 $T$ で得られる入力 $\hat{x}$ を $\hat{x}(T)$ とし、その所属クラスを $\omega_{q_0}(T)$ とすると各ジェスチャの生起確率 $P_T(G_j | \omega_{q_0}(T))$ は式(14)で計算され、これが最大となる $G_j$ を時刻 $T$ において最も行なわれている確率が高いジェスチャ $G(T)$ として選択する。

$$\omega_{q_0}(T) = \arg \min_{\omega_i} (D(\hat{x}(T), m_i)) \quad (13)$$

$$P_T(G_j | \omega_{q_0}(T)) = \frac{1}{K} \sum_{t=T-K+1}^T \frac{\sum \{x | x \in M_j, x \in \omega_{q_0}(T)\}}{\sum \{x | x \in \omega_{q_0}(T)\}} \quad (14)$$

$$G(T) = \arg \max_{G_j} (P_T(G_j | \omega_{q_0}(T))) \quad (15)$$

#### 3.4.4 姿勢推定

本稿における姿勢とは観測された入力 $\hat{x}$ (手の頭に対する3次元位置、速度)から各関節角度、角速度を表す20次元モデルベクトル空間上に1点 $\hat{\Theta}$ を決めることである。しかし前述の通り式(3)における $A$ はランクが不足するために解を一意に決定することはできない。そこで本稿では以下のように姿勢を推定する。

まずクラス $\omega_{q_0}$ に所属するジェスチャ $G_j$ のクラス代表点として各ジェスチャのクラス中心 $m_{q_0, G_j}, M_{q_0, G_j}$ とすると、 $\hat{\Theta}$ は式(16)で表せる。すなわち $B$ を求めれば姿勢を推定できることになる。

$$\hat{\Theta} = B \begin{pmatrix} y \\ 1 \end{pmatrix} + M_{q_0, G_j} \quad (16)$$

$$y = \hat{x} - m_{q_0, G_j} \quad (17)$$

次にクラス $\omega_{q_i}$ に所属する動きモデル $X$ と $\Theta$ に関しても同様に考えると式(18)が成り立つ。

$$\Theta = B \begin{pmatrix} Y \\ 1 \end{pmatrix} + M_{q_i, G_j} + e \quad (18)$$

$$Y = X - m_{q_i, G_j} \quad (19)$$

ここで十分に近傍のモデル同士ならば $\|e\|$ は十分に小さいと考えられ、 $X - m_{q_i, G_j}$ と $\Theta - M_{q_i, G_j}$ に相関があると考えられる。そこで重回帰分析により求められた $B$ による写像(式(16))を推定姿勢とする。以下にこれを用いた具体的な姿勢推定法を示す。

#### 姿勢推定法

1.  $i = 0$ , データの信頼性を判定する十分なデータ数をしきい値として $l=100$ とする。また、動きモデル $\Theta, X$ を保持する空のスタック $S_\Theta, S_X$ を用意する。ステップ2へ。
2. クラス $\omega_{q_i}$ において選択ジェスチャ $G_j$ の動きモデルを $S_\Theta, S_X$ に積む。またスタックのサイズが $l$ よりも大きければステップ3へ、小さければ $i = i + 1$ とし、ステップ2へ。
3.  $S_\Theta, S_X$ を用いて上記に示したように重回帰分析を行ない、 $B$ を求める。ステップ4へ。
4. 式(16)によって推定姿勢 $\hat{\Theta}$ を計算し、終了。

本手法では生起確率最大のジェスチャだけでなく、その他のジェスチャの姿勢も推定できる。生起確率が低くなるほど信頼できるデータ数を揃えるための領域が大きくなるので式(18)の $\|e\|$ は大きくなり、写像の近似も悪くなると考えられるが、全てのジェスチャについて姿勢推定を行なっておけば、実際に選択されなかったジェスチャが起こった場合でも選択ジェスチャの切替えによって安定した処理を行なうことができる。

#### 3.4.5 背景差分画像と形状モデルのマッチング

得られた推定姿勢 $\hat{\Theta}$ は誤差や選択ジェスチャの誤りが考えられるため、必ずしも入力画像に対して最適であるとは限らない。そこで $\hat{\Theta}$ を初期値として、背景との差分による人のシルエット領域と形状モデルの特徴点とのマッチングを行ない、各関節角度を最適化する。本稿では式(20)の評価関数 $C$ を最小化する $\Phi$ を求める。

$$C(\Phi) = - \sum_i S \cdot P \cdot f_i(\Phi) + d(\Phi) \quad (20)$$

$$d(\cdot) = \begin{cases} \infty & \text{if } \frac{x_{ei}^2}{l^2} + \frac{y_{ei}^2}{s^2} < 1, \\ 0 & \text{otherwise} \end{cases} \quad (21)$$

$f_i$  は人物モデルの各特徴点  $i$  (肘, 肩, 腕の各局所座標系上の点) を世界座標系に変換する関数,  $P$  は画像面上への射影変換,  $S$  は画像上に投影された特徴点位置の人物シルエット画像 (2 値画像) の画素値 (1 または 0) を返す関数である.  $d(\Phi)$  は上体を近似した楕円内に肘が存在しないようバイアスをかける関数である. 以下に具体的なアルゴリズムを示す.

#### モデルマッチング

1. 全て ( $J$  個) のジェスチャの推定姿勢  $\hat{\Theta}_j (1 \leq j \leq J)$  を探索初期点として姿勢候補行列  $\phi$  に入れ, ステップ 2 へ.

$$\phi = (\phi_1, \phi_2, \dots, \phi_J) \quad (22)$$

$$\phi_k = \hat{\Theta}_k \quad (1 \leq k \leq J) \quad (23)$$

2. 式 (14) で計算された  $J$  個の生起確率を比率としてランダムに一つのジェスチャ  $K$  を選択し, 姿勢候補  $\Phi = \phi_K$  とする. ステップ 3 へ.
3.  $C$  が最小, または規定のループ回数を越えたら終了. そうでなければ平均 0, 分散  $\sigma$  の正規分布関数  $N(0, \sigma)$  によって式 (24) を用いて  $\phi_k$  を更新してステップ 2 へ.

$$\phi_k = \phi_k + N(0, \sigma) \quad (24)$$

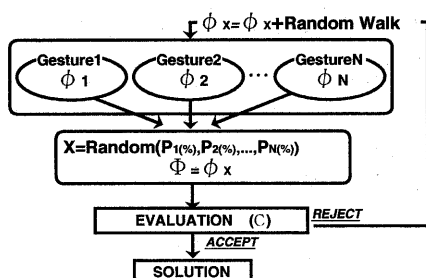


図 4: 姿勢の最適化

#### 3.4.6 姿勢予測

一般に画像処理は大きなコストを要する処理で, 予め姿勢を予測できれば様々な効率化が期待できる. また, ノイズによって信頼できるデータが得られなかった場合には過去の予測姿勢で代替することによって安定に推定を行なえる. 本手法では, 動きモデルを

表 3: 推定誤差

| -                      | Shoulder | Elbow  | Hand   |
|------------------------|----------|--------|--------|
| Mean(mm)               | 26.4     | 48.5   | 48.9   |
| Var.(mm <sup>2</sup> ) | 15.6     | 1356.3 | 3574.4 |

利用してジェスチャに適した角速度が計算できるので, これを用いて次の時刻での関節角度を予測し, 順運動学により 3 次元位置を計算する.

### 3.5 実験と考察

#### 3.5.1 ジェスチャの生起確率

入力として「パンチ」→「何もしない」→「バイバイ」を実行するシーケンスを用いて各ジェスチャの生起確率を調べた. 生起確率は式 (14) の  $K=5$  とし, 過去 5 フレームの確率を平滑化して算出して最終的な確率とした. 図 5 に結果を示す.

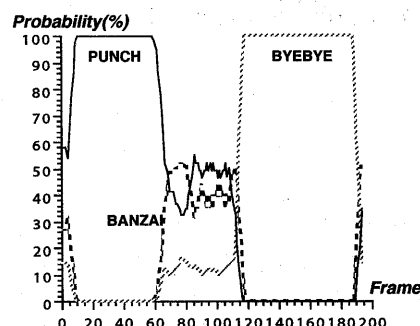


図 5: 生起確率の遷移

結果をみるとジェスチャに応じて理想的に遷移していることが分かる. 何もしていない状態のときにはバンザイとパンチの生起確率が同程度になっているが, 姿勢に大きな違いはないので姿勢推定には影響がない. またバンザイの画像を入力としたシーケンスでも実験を行なったところ同様の結果が得られた.

#### 3.5.2 動きモデルによる姿勢推定

図 5 と同じ入力画像シーケンスを用いて動きモデルによる姿勢推定を行なった. 推定した姿勢は生起確率が最も高いジェスチャについてのみ行ない, 評価は右腕に関して, 実際の手, 肘, 肩の位置と推定姿勢との誤差を距離で計測した. また実際の値は手で入力することによって計算した. 結果を図 6 と表 3 に示す.

結果を見るとどの関節も平均で 50(mm) に収まっていることが分かる. 図 6 を見てもほぼ安定して姿

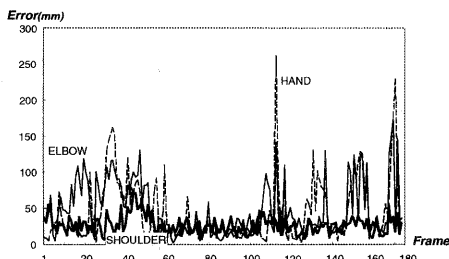


図 6: 右手, 肘, 肩の推定値と手で与えた値との誤差

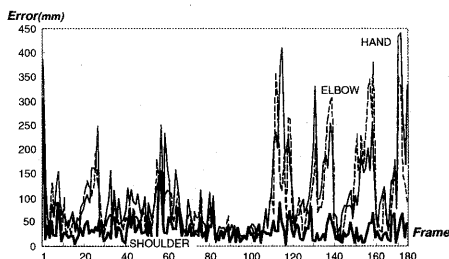


図 7: 右手と右肘の予測値と手で与えた値との誤差

勢を推定できているので, 本手法を使って姿勢を最適化探索の初期値とする方法は有効である. また 116 フレーム目で手の推定が悪くなっている原因としては「何もしていない」→「バイバイ」のジェスチャの切れ目であることが考えられる. 本手法では過去 5 フレームの生起確率を平滑化して最終的な生起確率を算出しているため, ジェスチャの切れ目では立ち上がり時に式 (14) で示したように  $K=5$  フレーム必要になる. これはモデルへの追従度の問題であり,  $K$  を小さくしすぎるとノイズに影響を受け, 大きいとジェスチャの変化への追従度が遅くなってしまう. また, 手の頭に対するの 3 次元相対位置が入力されてから動きモデルによる姿勢を推定するまでの実行時間を計測したところ平均 10.2(msec) であった.

### 3.5.3 動きモデルによる姿勢予測

図 5 と同じ入力画像シーケンスを用いて動きモデルによる姿勢予測を行なった. 評価は右腕に関して, 手と肘の位置の真値と姿勢予測との誤差を距離で計測した. 結果を図 7 に示す.

図 7 を見ると, 動きの比較的激しい「バイバイ」では大きく誤差が出ているが, 手の予測誤差平均は 106.8mm であった. また, 「バンザイ」を入力画像に用いて各関節に関して

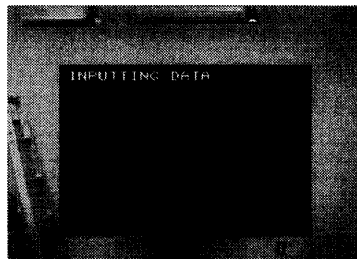


図 8: フロー抽出の例

1. 前フレームの速度は保たれると仮定した線形予測
2. 静止している仮定した予測
3. 動きモデルによる予測

の 3 つについて同様の実験を行なった. 結果を表 4 に示す.

表 4: 予測誤差

|          | 1            |                            | 2            |                            | 3            |                            |
|----------|--------------|----------------------------|--------------|----------------------------|--------------|----------------------------|
|          | Mean<br>(mm) | Var.<br>(mm <sup>2</sup> ) | Mean<br>(mm) | Var.<br>(mm <sup>2</sup> ) | Mean<br>(mm) | Var.<br>(mm <sup>2</sup> ) |
| Shoulder | 30.5         | 388.5                      | 28.5         | 99.1                       | 24.9         | 212.0                      |
| Elbow    | 73.0         | 924.3                      | 63.0         | 1048.4                     | 57.8         | 1169.7                     |
| Hand     | 55.9         | 1617.1                     | 47.0         | 1028.0                     | 43.4         | 1170.0                     |

結果を見ると全ての関節で動きモデルを用いた場合が最も良い予測を与えていることが分かる. よって動きモデルを用いた姿勢予測は有効といえる.

## 4 動きを利用したリアルタイムジェスチャ認識

ここでは, 動き情報を利用してジェスチャを認識する手法を紹介する. 動き情報を利用するには人物を検出して動き情報の切り出しをする必要がある. ここでは, 人物領域の切り出しが完了し, 動き領域の切り出しができるとして話を進める.

### 4.1 動きベクトルの計算

図 8 にフローの抽出例を示す. 動きベクトルは大量のデータを発生するので, KL 展開を用いて入力データの圧縮を行なう. ジェスチャは圧縮された部分空間上である軌道を描くので, モデル軌道とのマッチングを行えば良い.

### 4.2 HMM によるジェスチャの認識

HMM による認識の利点は, 入力データの時間変動に対してロバストである点で, また, 多人数に対

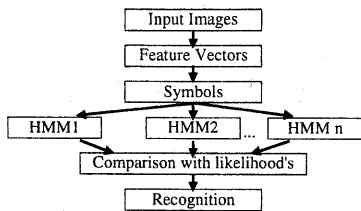


図 9: HMM を用いた認識のフレームワーク

しても学習により認識率を向上できる点である。ただし、この利点は欠点でもあり、認識率を高めるための学習データはかなりの量が必要である。

図 9 に HMM を用いた認識のフレームワークを示す [7]。本手法は、カメラから得られる動画像において人物の動き情報に注目し、それをテンプレートマッチング法で動き領域を抽出することで、2次元のモーションベクトルを求める。それぞれのジェスチャにおいて各フレームごとにモーションベクトルを計算し KL 展開することでジェスチャ空間を作成する。また、HMM をジェスチャ認識に応用するためには、動画像をシンボルの時系列に変換する必要がある。そこで、ジェスチャ空間に投影した各主成分に対して、それらを量子化することでシンボル列に変換する。変換されたシンボルの時系列を HMM で学習させ、各 HMM 内での尤度を計算し、最大となる HMM のモデルを選択することで認識を行なう。

KL 展開で得られた部分空間のクラスタリングには、VQ (ベクトル量子化) 法で行なう。代表点はベクトルの統計的分布を反映するように選ばなければならないが、この解析的手法は知られておらず、LBG クラスタリングアルゴリズム [8] と呼ばれる繰り返しアルゴリズムによって作成する。

#### 4.3 オートマトンを用いた HMM の階層化

本システムでは、カメラから得られた画像からオプティカルフローを特徴量として抽出し、それをベクトル量子化することでシンボルを生成し、そのシンボル系列と HMM とのマッチングの出力確率を Viterbi アルゴリズムで計算するまでの過程を実時間で実現することができる。しかし、このようなシステムではどこからどこまでが 1 つのジェスチャのシーケンスであるかを切り出すことが重要となる。従来の離散 HMM とベクトル量子化の技術だけではフレームごとの認識はできても、1 つのジェスチャに対応するシーケンスの切り出しはできない。そこで、本手法では各フレームでの HMM と観測系列のマッ

ングによる出力確率をオートマトンを通して階層化することにより望みのシーケンスを切り出す方法を提案する。

HMM  $\lambda_G$  の出力確率密度分布  $p_G(y)$  は正規分布に従うものとする、解決器では

$$A(y, G) = P(\lambda_G)P(y|\lambda_G) \quad (25)$$

$$= k_G \sum_{t=1}^{\min(N, T-T_R)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(p_G(y_{T-t+1})-\mu)^2}{2\sigma^2}}$$

をもとめ、 $A(y, G)$  がヒューリスティックに決めたあるしきい値を越えたときジェスチャー  $G$  を出力結果とする。ここで、 $k_G$  は観測フレーム中の  $G$  の出現数で、事前確率  $P(\lambda_G)$  の近似として用いる。また、 $T_R$  をなんらかのジェスチャーを認識した時刻とする。

本手法は内部状態を 2 つ持つオートマトンを用いたモデルで表せる。オートマトンの入力は  $A(y, G)$  がしきい値以上ならば 1、しきい値以下ならば 0 である。また出力は、ジェスチャーのシーケンスを  $G$  と認識したとき 1 で、そうでないとき 0 である。初期状態を  $S_0$  とし、 $A(y, G)$  がしきい値をこえない間はジェスチャー  $G$  と判断せずに状態  $S_0$  に留まる。 $A(y, G)$  がしきい値を越えたとき入力系列をジェスチャー  $G$  と認識して状態  $S_1$  に遷移し、 $T_R$  を設定する。そしてまた状態  $S_0$  へと遷移する。

#### 4.4 実験と考察

「あけまして」、「おめでとう」、「こんにちは」、「お久しぶり」、「お元気で」、「さようなら」、「ご苦労様」の 7 つの手話ジェスチャに、さらにシステムのジェスチャ認識窓への出入りパターンを加えた 9 つの HMM を学習により作成し、フレーム毎の認識実験を行なった。

学習データは 1 名  $\times$  50 フレーム  $\times$  50 個用い、認識には 5 名  $\times$  100 フレームのデータを用いた。以下にそれぞれのジェスチャの認識率を示す。

「お元気で」のジェスチャの認識率が悪いのは、「こんにちは」とジェスチャの最後の部分が似ており混同がおきているためである。また、「出」のパターンにはさまざまな種類があり正しく HMM が学習されていないことがわかる。その他のジェスチャについては概ね学習がうまくいっており高い認識率を維持している。

#### 5 おわりに

統計的動きモデルを用いることによってジェスチャー動作の姿勢を効率良く推定する方法を提案した。

表 5: ジェスチャ認識率 (%)

| ジェスチャ | あけ    | おめ   | こん   | おひ   | おげ   | さよ   | ごく   | 入     | 出    |
|-------|-------|------|------|------|------|------|------|-------|------|
| あけまして | 100.0 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0   | 0.0  |
| おめでとう | 6.9   | 93.1 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0   | 0.0  |
| こんにちは | 10.7  | 0.0  | 89.3 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0   | 0.0  |
| お久しぶり | 4.9   | 0.0  | 0.0  | 95.1 | 0.0  | 0.0  | 0.0  | 0.0   | 0.0  |
| お元気で  | 24.4  | 0.3  | 41.8 | 0.0  | 33.5 | 0.0  | 0.0  | 0.0   | 0.0  |
| さようなら | 8.9   | 0.0  | 0.0  | 0.0  | 0.0  | 91.1 | 0.0  | 0.0   | 0.0  |
| ご苦労様  | 10.7  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 89.3 | 0.0   | 0.0  |
| 入     | 0.0   | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 100.0 | 0.0  |
| 出     | 13.7  | 0.0  | 0.0  | 0.0  | 0.0  | 20.0 | 17.9 | 19.3  | 29.1 |

時系列ステレオ画像から色と位置を利用した領域抽出を行って頭と手の領域を抽出し、ステレオマッチングを行なうことで対象の3次元位置を求めた。頭と手の3次元位置と形状モデルを利用し、統計的な動きモデルを用いることによって各ジェスチャの生起確率を計算した。さらに各ジェスチャに関して姿勢を推定し、シルエット領域とのモデルマッチングによる最適化を行なう際の探索初期点とした。実画像に対して実験を行ない、提案手法の有効性を示した。本手法はジェスチャを対象としたが、予め動きの分かっているような場合にはモデルを作ることによって様々な対象を扱うことができる。

また、HMMとオートマトンを階層的に組み合わせて、ジェスチャを認識する方法を提案した。そして、フレーム毎の認識率を手話ジェスチャーにより評価し概ね学習がうまくいっていることを示した。

今後の課題としては姿勢の最適化を実装し、実験することが考えられる。またジェスチャの種類を増やしたり、欠損値が生じたときの対処などを実装する必要もある。

## 謝辞

本研究の一部は、日本学術振興会未来開拓学術研究推進事業 (JSPS-RFTF 99P01404) の補助を受けた。

## 参考文献

- [1] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima: "Real-Time Estimation of Human Body Posture from Monocular Thermal Images"; *In Proc. of CVPR*, pp. 15-20 (June 1997)
- [2] C. Wren, A. Azarbayejani, T. Trevor Darrell, and A. Pentland: "Pfinder: Real-Time Tracking of the Human Body"; *In Proc. 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 51-56 (1996)
- [3] 島田伸敏, 白井良明, 久野義徳, 三浦純: 「形状の個体差に適應する精密なジェスチャ推定」; *インタラクシオン*, pp. 5-12 (1998)
- [4] 島田伸敏, 白井良明, 久野義徳: 「三次元モデルを用いた二次元動画画像からの手指姿勢の推定」 *信学技報 PRU*, pp. 25-32 (May 1994)
- [5] Ming-Hsuan Yang and N. Ahuja: "Extracting Gestural Trajectories"; *In Proc. International Conference on Automatic Face and Gesture Recognition*, pp. 10-15 (1998)
- [6] 大垣健一, 岩井儀雄, 谷内田正彦: 「動きと形状モデルによる人物の姿勢推定」; *信学論*, Vol. J82-D-II, No. 10, pp. 1-11 (Oct 1999)
- [7] 畠直志, 岩井儀雄, 谷内田正彦: 動き情報と情報圧縮を用いたロバストなジェスチャ認識手法. *電子情報通信学会誌*, J81-DII(9):1983-1992, Sep. 1998.
- [8] Yoseph Linda, Andres Buzo and Rovert M. Gray. An algorithm for Vector Quantizer design, *Computer*, 28, pp. 84-94, 1980.
- [9] J. Yamato, J. Oya, and K. Ishii. Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model, *In Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition*, pp. 379-385, 1992.
- [10] M. Schuster and G. Rigoll. Fast Online Video Image Sequence Recognition with Statistical Methods, *In Proc. IEEE Inc. Conference on Acoustics, Speech and Signal Processing*, pp. 3450-3453, Atlanta, 1996.
- [11] S. Kuo and O. Agazzi. Keyword Spotting in Poorly Printed Documents Using Pseudo 2-D Hidden Markov Models, *IEEE Trans. on Pattern Recognition and Machine Intelligence*, Vol.16, No.8, pp. 842-848, 1994.
- [12] S. Eickeler, S. Müller, and G. Rigoll. High Quality Face Recognition in JPEG Compressed Images, *In Reoc.IEEE Internal Conference on Image Processing*, Kobe, 1999.
- [13] S. Müller, S. Eickeler, and G. Rigoll. Crane Gesture Recognition Using Pseudo 3-D Hidden Markov Models" *Forth IEEE International. Conference on Automatic Face and Gesture Recognition*, 2000, pp.398-402.