Vol.15 No.1, 2013

原著論文

HMMを利用した画像処理による手話単語の認識のための特徴抽出および状態分割

松尾 直志 *1　山田 寛*2　白井 良明*3　島田 伸敬*3

Feature Extraction and State Decomposition for Image-based Japanese Sign Language Words Recognition Using HMM

Tadashi Matsuo*1, Yutaka Yamada*2, Yoshiaki Shirai*3 and Nobutaka Shimada*3

Abstract – **This paper proposes a method of extracting efficient features for recognizing Japanese sign language words and recognizing them with Hidden Markov Model. Hand and face regions are extracted and tracked by image processing technique. Features of hand motion and shape are calculated by the extracted regions. Each HMM model is trained by Baum-Welch algorithm. The number of states is determined based on segmentation of the hand motion. Viterbi algorithm is employed for recognition. Experimental results show the validity of the proposed method.**

Keywords : 手話認識，画像処理，隠れマルコフモデル

1. はじめに

聴覚障害者が意思を伝えたい場合，筆談に比べると手話は格段に便利である．また，手話通訳者の不足などのためにも，手話通訳システムが望まれている．

手話の認識は，時間的に変化する手話の特徴を抽出する処理と，抽出された特徴系列から手話を認識する処理からなる．手話は，ジェスチャーと比較すると，手指の形（手形）の種類が多い，手形や位置が高速に変化する，同じ色の両手同士や顔との隠蔽があるという性質があるため，特徴を抽出することが困難である．

特徴を抽出する手法は，特殊な手袋を利用する手法と，画像を利用する手法に大別される．前者の典型例であるデータグローブを用いる手法 [1]~[3] は，高速で信頼性の高い特徴抽出が可能であるが，特殊な設備が必要であり，話者への負担を要する．

一方，画像処理を利用した特徴量の取得は話者への負担は少ないが，特徴抽出が難しい．そこで，特別な色をつけた普通の手袋を装着し手話を行い，その画像系列から簡単な手話を認識する研究も行われた [4] が，手話者への負担がある．また，赤外線カメラを用いた温度画像を用いた手法 [5] もあるが，安価な可視光のカメラを使えないという制限がある．

特別な器具を装着させず，手話から画像処理によって特徴抽出を行う研究では，オプティカルフローを利用した手法 [6], [7] や動きベクトルを用いた手法 [8]，フレーム間差分を利用した手法 [9] があるが，これらの手法は大局的な動きを検出することはできるが，手の形状については十分考慮されていない．また，手や顔との隠蔽への対応も不十分である．

今川ら [10] は手の見えを特徴として認識に利用しているが，顔と手の重複時の手領域の抽出のためには布で顔を隠すなど不自然な条件を用いて学習を行っている．山本ら [11] や柳ら [12] は手領域の重心と主軸方向を手の形状特徴としているが，手や顔との隠蔽への対応が不十分である．

画像からの手話認識は，現在，米国，フランス，台湾，日本などで研究されているが [14]，手と手や手と顔の隠蔽状態における特徴抽出の困難さと，処理速度が遅い，学習用の手話画像が簡単に得られないなどの理由で，まだ実用には至っていない．

特徴系列から手話を認識する手法に関しては，DPマッチング [6]，自己組織化マップ (SOM) [3], [9] 等もあるが，隠れマルコフモデル (Hidden Markov Model, HMM) が提案されて以来 [13]，時間の伸縮を許容するHMMが使われることが多い [11], [12]．HMMは，音声認識のように，ある程度の変動をともなう時系列パターンをモデルに対する尤度によって識別する場合に適した手法であるので，本論文でも特徴系列の認識には，HMMを採用する．HMMを用いた認識には予め単語モデルを作成する必要がある．単語モデルを学習する際，初期に，HMMの状態数，状態遷移，各状態の出力ベクトルの分布など（初期モデルという）が必要となる．従来研究では，手動で状態構造と初期値を

*1: 立命館大学 総合理工学研究機構
*2: 立命館大学大学院 理工学研究科
*3: 立命館大学 情報理工学部
*1: Research Organization of Science and Engineering, Ritsumeikan University
*2: Graduate School of Science and Engineering, Ritsumeikan University
*3: School of Information Science and Engineering, Ritsumeikan University

与え初期モデルを作成する場合，もしくは全ての単語で共通の状態数を与え作成する場合のいずれかであった．しかし，手話単語は膨大なため手動で全ての単語の初期モデルを作成するのは困難である．また，共通の状態数では単語ごとの複雑さの違いが考慮できない．

本論文では，画像からの手話認識で困難な課題である手話特徴の抽出および HMM による単語モデル作成のための特徴系列の状態分割を扱う．すなわち，色情報と動きの拘束を用いて，手領域と顔領域を抽出し，手話特徴を求める．手同士あるいは手と顔の隠蔽が発生した場合は，隠蔽の間は隠蔽直前の手領域と直後の手領域のいずれかに類似しているという仮定のもとで，手の位置と方向を決定する．

手話特徴量は，左右それぞれの手の位置に関する特徴として，画像上での手の位置と速度，手の形状に関する特徴として，画像上での手の方向と伸ばしている指の数を用いる．位置については，顔付近では正確な位置が必要であるが，顔から離れればある程度の位置の違いは許されるべきと考えられるので，位置と速度は顔を原点とする極座標系で表す．

Baum-Welch アルゴリズムによる HMM の学習では，あらかじめ HMM の状態数と状態遷移構造を定めておく必要がある．本論文では遷移構造を状態が直線状に接続されたものに限定し，単語ごとに典型的な特徴時系列データを状態分割して状態数を決定する．手話は手の構え，向き，提示位置，大局的な運動からなる形態素の時系列の連続としての表現で語の意味が主に決まる [17] ので，ここでは手の動きに注目して状態分割を行う．HMM を学習した後，手話の認識には Viterbi アルゴリズムを用いる．

2 章では，画像系列から手，顔領域を抽出する手法について述べ，3 章では，用いる手話特徴量の定義とその求め方について述べる．4 章では認識に利用する HMM の学習と認識を簡単に述べ，特徴系列の初期状態分割を説明する．5 章では，認識実験と考察を述べ，本手法の有効性について検討する．

2. 手，顔領域の抽出

まず背景領域と手話者（人物という）領域に分け，人物領域の中から肌色領域を抽出する．肌色領域を用いて，左右の手と顔領域を抽出し，時系列にわたって各領域を追跡しながら領域を抽出する．

2.1 人物領域の抽出

画像から人物領域を抽出するため，背景のみの画像と比較して，色や輝度（以後属性と言う）の差が一定以上の領域を人物領域として抽出する．この場合，属性をどのように表現して差を求めるかが問題となる．

普通の RGB 表色系では各成分に色情報と輝度情報が含まれる．したがって，輝度の影響を強く受け，類似の色でも属性の差が大きくなりすぎる．

これに対して，色と輝度値を分離した成分で表す HSV 表色系が提案されている．この表色系では，V（輝度）の重みを変えることにより照明の変化に対応できる利点がある．しかし，S(彩度) が低い場合は，H(色相) の値は信頼性が低く，V(明度) が低い場合は H，S ともに信頼性が低い．したがって，HSV 表色系で属性の差を求めることは適当でない．

そこで，本研究では影は影のない領域と比べて輝度値が低くなり，S,V が低い場合は H の差の影響を小さくする新しい表色系を用いる．この表色系 (ξ, η, ζ) は，HSV 表色系を以下の式を用いて変換する．

$$\begin{cases} \xi = S \times (V/100) \times \cos(H) \\ \eta = S \times (V/100) \times \sin(H) \\ \zeta = V \times \omega \end{cases}$$

ここで，$0 \leqq H < 2\pi$, $0 \leqq S \leqq 100$, $0 \leqq V \leqq 100$, ω は色の差に対する輝度の差の重みを表し，画像の明るさや衣服の色などに依存し，影領域の誤抽出を出来るだけ防ぎ，人領域を正しく抽出できるように設定する（ここでは 0.5 とした）．この ($\xi\eta\zeta$ 表色系) は，V が小さい時は $\xi\eta\zeta$ の値がすべて小さくなり，S 値が小さい場合は $\xi\eta$ の値が小さくなる円錐形の表色系である．$\xi\eta\zeta$ 表色系で背景差分によって人物領域を求める．

対象画像の画素の属性 $(\xi\eta\zeta)$ と背景画像の属性 $(\xi_b\eta_b\zeta_b)$ の差 d は次式で表す．

$$d = \sqrt{(\xi - \xi_b)^2 + (\eta - \eta_b)^2 + (\zeta - \zeta_b)^2} \qquad (1)$$

対象画像の座標 (x, y) における d を $d(x, y)$ として，人物候補領域 H_c を次式で表す．

$$H_c = \{(x, y) \quad | d(x, y) \ > \ d_t\} \qquad (2)$$

d_t は (x, y) が人物候補領域か背景領域かを判断する閾値で，ここでは実験的に 30 としている．H_c の内，連結領域の面積が最大のものを人物領域とする．

2.2 肌色領域の抽出

肌の色は個人や照明条件によって差があるため固定の閾値で肌色を定義するのは難しい．そこで動画像の初期フレームの肌の色から，肌色を定義する [16]．

初期フレームの画像から手動で肌の画素をサンプリングし，$\xi\eta\zeta$ 表色系の $\xi\eta$ 空間における肌色の分布が正規分布であると仮定し，その 90%等確率楕円内の色を肌色とする．(確率値が変わると，肌色領域がやや収縮拡大する)．図 1 に例を示す．

2.3 手，顔領域の抽出

得られた肌色領域についてそれぞれどの領域が手，あるいは顔の領域かを決定する必要がある．

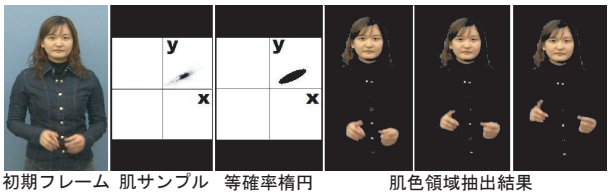


図 1　肌色抽出
Fig. 1　Extraction of skin regions.

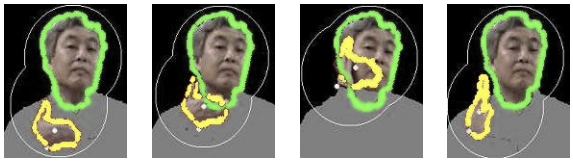


図 2　バックトラック無の場合の手領域抽出
Fig. 2　Hand extraction by only forward tracking.

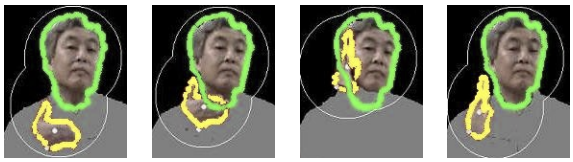


図 3　バックトラック有の場合の手領域抽出
Fig. 3　Hand extraction by forward and backward tracking.

得られた領域の内，一定以上の面積を持つもので，面積が大きいものから順に最大 3 つの領域を得る．それ以外はノイズとみなし除去する．

我々は初期フレームでは手が腰の辺りにあるという仮定をおいている．初期フレームではその仮定を利用して顔領域，手領域を決定する．

以降のフレームでは，前フレームでの位置を基にそれぞれの領域の追跡を行う．顔領域は手領域に比べて手話中にあまり動かないので，顔領域は前フレームでの顔領域に最も近い領域を顔領域とする．手の領域は等速運動を仮定した際の予測位置に最も近い領域をそれぞれの手の領域とする．

2.4　隠蔽時の手, 顔領域の抽出

手話の際には手と顔や手同士の隠蔽がしばしば生じる．重なった領域において正確な手や顔領域の抽出は困難であるため，隠蔽が起こる直前の手や顔の画像をテンプレートとして保存し，探索領域 A 内で回転，平行移動をしながらテンプレートマッチングを行う．しかし，手話の際には図 2(b)(c) のように隠蔽中に手の形が変わりテンプレートが実際の手の形に合わない場合がある．また，手と顔の隠蔽が生じた場合は，顔領域が手に比べて大きいため，手が顔の中に完全に入ってしまう場合がある．

そこで，隠蔽が終了した直後の手や顔の画像をテンプレートとしてバックトラックを行う．バックトラックを追加した場合の結果を図 3 に示す．図 3(c) は隠蔽直前の図 3(a) とは形が変わっているが，隠蔽直後の図 3(d) の手領域をテンプレートとしてバックトラックを行うことで抽出結果が改善できる．

2.4.1　手同士の隠蔽の場合

手同士の隠蔽が生じた場合はテンプレートマッチングを行い，手の位置を検出する．なお，隠蔽中に手の大きさはあまり変わらないとし，テンプレートのスケールの変換は行わない．テンプレートマッチングの (x_0, y_0) における評価式 $S_d(x_0, y_0)$ は下式とする．

$$\begin{aligned} S_d(x_0, y_0) = & \\ \sum_{(x,y)\in T} & \left(\{\xi_T(x,y) - \xi(x+x_0, y+y_0)\}^2\right. \\ & + \{\eta_T(x,y) - \eta(x+x_0, y+y_0)\}^2 \\ & \left. + \{\zeta_T(x,y) - \zeta(x+x_0, y+y_0)\}^2\right) \end{aligned} \tag{3}$$

ここで，T はテンプレートの領域，ξ, η, ζ は入力，ξ_T, η_T, ζ_T はテンプレートの $\xi\eta\zeta$ 色成分である．重心位置 $(\hat{x}, \hat{y})$ を下式により決定する．

$$(\hat{x}, \hat{y}) = \arg \min_{(x_0,y_0)\in A} S_d(x_0, y_0) \tag{4}$$

ここで，A は探索領域であり，手の運動を等速運動で近似し，前フレームの手の位置からの予測位置の近傍とする．A は下式により決定する．

$$A = \{(x,y) | \sqrt{(x-x_p)^2 + (y-y_p)^2} \leqq r_m + r(v)\} \tag{5}$$

ここで，(x_p, y_p) は手の予測位置，r_m は探索範囲の最低半径を表す定数，$r(v)$ は手の運動の速さに比例する．

隠蔽領域中でどちらの手が手前にあるかは手の評価値の合計 S_{sd} によって判断する．S_{sd} を下式に示す．

$$S_{sd} = S_d^f(\hat{x}_p, \hat{y}_p) + S_d^s(\hat{x}_q, \hat{y}_q) \tag{6}$$

ここで添字の S_d^f は先にマッチングする側，S_d^s は後にマッチングする側を表し，p, q はそれぞれの手を表す．先にマッチングしたテンプレートと重なる肌色領域は，他方の手領域ではないため黒色 $(\xi = \eta = \zeta = 0)$ として $S_d^s(\hat{x}_q, \hat{y}_q)$ を計算する．S_{sd} を $p =$ 右手，$q =$ 左手とした場合と逆の場合と二通り計算し，小さい方の結果をマッチング結果とする．その時の p が手前にあるとする．

順方向のマッチング終了直後の領域をテンプレートとして逆方向へバックトラックを行う．t フレーム目の S_{sd} である $S_{\mathrm{sd}}(t)$ を順方向とバックトラックそれぞれ計算し，バックトラックを続けるか否かを判断する．下式の条件を満たした時にバックトラックを終了する．

$$S_{\mathrm{sd}}^f(t) < S_{\mathrm{sd}}^b(t) \tag{7}$$

ここで，添字の f は順方向，b は逆方向を表す．

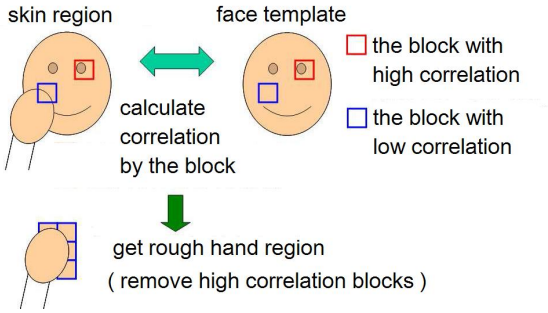
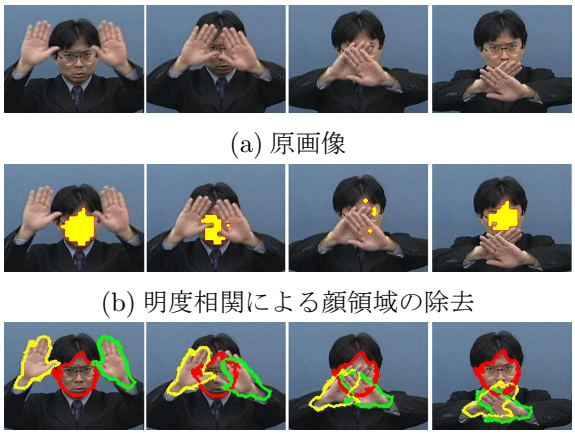


図 4　相関を利用した顔領域の除去
Fig. 4　Face removal by luminance correlation.



(a) 原画像

(b) 明度相関による顔領域の除去

(c) マッチング結果

図 5　顔と手領域の分離
Fig. 5　Example of face-hand separation.

2.4.2　手と顔の隠蔽の場合

手と顔の隠蔽が生じた場合は，顔の位置はあまり変化がないという仮定のもと，先に顔のテンプレート位置を決定する．図 4 に示すように，顔テンプレートとの明度相関が高い領域を顔領域として A から除外した後，マッチング処理を行う．

まず式 (3)，(4) によって顔テンプレートの位置を決定する．次に，顔テンプレート及び，肌色領域を 5×5[pixel] のブロックに分割する．ブロック毎に明度相関を求め，相関値が 0.5 以上の領域を顔領域として A から除外する．相関値の閾値は，顔の向きや表情の変化，および種々の手の形の画像がどれくらい顔画像と類似しているかに依存する．値が大きすぎると手の探索に時間がかかり，小さすぎると顔以外を除外するので，ここではやや高めにしてある．図 4 のように，A から大まかな顔領域を除外できる．残りの A について式 (3)，式 (4) を用いて手の位置を検出する．

両手と顔が隠蔽する場合は，まず顔テンプレートの位置を決定し，相関による顔領域の除外を行う．その後は，手同士の隠蔽と同様にテンプレートマッチング処理を行う．図 5 に相関による顔領域の除外とそのあとのテンプレートマッチングの例を示す．

3.　手話特徴

神田ら [17] によると，手話単語は手のかまえ，向き，提示位置，大局的な運動を形態素としている．それぞれの特徴の計算方法について述べる．

3.1　手の位置に関する特徴

画像処理による手話認識では XY 座標系による手の位置や速度，加速度を特徴とするものが多い [9], [11], [12]．しかし，顔の一部を指差す単語では細かな位置の違いが重要な意味を持つため精確な位置が必要となる．一方，顔から離れた位置では手の位置にばらつきがあり位置の小さな違いは許容する必要がある．そこで本研究では，顔の位置を基準とする対数極座標系 (r, θ) で手の位置を表現する．

まず，手の重心と顔の重心の間のユークリッド距離 r'_r, r'_l を求める．添字の r, l は右手，左手を表す．初期フレームでは手が腰付近にあるという仮定のもと，初期フレームの r'_r, r'_l の平均 r'_{ave} で下式のように長さを正規化する．

$$r''_r = r'_r / (r'_{\mathrm{ave}}) \tag{8}$$

ここで，r''_r は正規化された r'_r とする．求めた r''_r を下式によって対数 r_r へ変換する．

$$r_r = \ln (r''_r + 1) \tag{9}$$

左手についても同様にして r_l を求める．

θ_r, θ_l は顔を基準とした右手 (左手) の方位角である．x 軸の正の方向と，顔 - 右手 (左手) の重心座標を結んだ直線のなす角とする．

(r, θ) は顔と手の位置関係を表す特徴であるが，手と手の相対的な位置関係が重要な場合もある．そこで，左手からみた右手の相対座標 (x, y) も特徴とする．上述の (r, θ)，(x, y) を用いて次に示す 10 種類の量を手の位置に関する特徴として利用する．

- 右手 (左手) と顔の重心間の距離の対数 $r_r, (r_l)$
- 前フレームの $r_r, (r_l)$ との差分 $\Delta r_r, (\Delta r_l)$
- 右手 (左手) の顔からの方位角 $\theta_r, (\theta_l)$
- 前フレームの $\theta_r, (\theta_l)$ との差分 $\Delta\theta_r (\Delta\theta_l)$
- 左手からみた右手の相対座標 (x, y)

3.2　手の形状に関する特徴

図 6 のように動きが似ていても手の形状が違うと意味が異なることがあるため，手の形状に関する特徴も認識に必要となる．手の形状に関する特徴は以下の 4 つの量を用いる．

- 右手 (左手) の突起数 $N_r (N_l)$
- 右手 (左手) の向き $\{u(1-r), v(1-r)\}_r$, $(\{u(1-r), v(1-r)\}_l)$

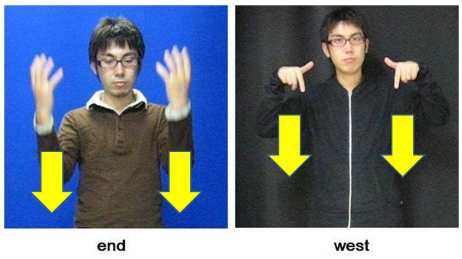


図 6　動きが似ているが手の形状が異なる単語
Fig. 6　Words that have similar motion but have different hand shape.

3.2.1　突起数

突起数は画像から確認できる伸ばしている指の数である．図 7(a) のように手首点から手の輪郭点の距離を順序づけ，第 s 点での手首点との距離 $l(s)$ とする (図 7(b))．突起数は $l(s)$ に基づいて突起数を決定する．手首点は肘点に最も近い手の輪郭点とする．

まず肘点は，人物領域の輪郭線を距離変換した画像と肘テンプレートのマッチングにより決定する [16]．一般に手話は胸元より上で行われるため，円弧を肘テンプレートとして用いる．肘点の座標 (x_e, y_e) は下式により決定する．

$$(x_e, y_e) = \arg \min_{(x_0,y_0)\in A} \sum_{(x,y)\in T} I(x+x_0, y+y_0) \quad (10)$$

ここで，T は肘テンプレートの領域，$I(x,y)$ は，距離画像の座標 (x,y) における画素値，A は探索領域である．A は画像全体とする．図 8 に肘点の検出例を示す．

次に，$l(s)$ に基づいて突起数を求める．まず，$l(s)$ の極大値をとる点を山点，極小値をとる点を谷点とする．しかし，山点，谷点はノイズの影響により多数出る．そこで，以下の条件を満たす山点，谷点をノイズとして除去する．

1. 隣接する谷点との差が閾値 d_d^c 以下の山点
2. 1 を挟む谷点の内大きい値をとる方
3. $l(s)$ が閾値 d_m^c 以下の山点
4. 3 を挟む谷点の内大きい値をとる方

上述の処理を除去される点がなくなるまで繰り返し，残った山点の数を突起数とする．極大値の両隣の極小値をとる輪郭点が閾値 n_w^c 点以上離れている場合は握りこぶしであったと判断して突起数は 0 とする．それぞれの閾値は，正面に向けて開いた手の長軸の長さ d_p を基準として，$d_m^c = 0.5d_p$，$d_d^c = 0.1d_p$，$n_w^c = d_p$ とした．閾値の変動によって，突起数は変化する場合もあるが，学習と認識で同じ閾値を用いる限り，認識率に大きな影響はない．

3.2.2　手の方向

手の方向は，手領域の慣性主軸の方向とする．方向は 2π の周期性があるので，連続する方向が不連続となる．また，手領域が円に近い場合は慣性主軸方向の信頼性がない．すなわち，円に近いほど方向の差が特徴の差に影響を与えないことが望ましい．そこで，手を楕円近似した際の長軸 a_l と短軸 a_s の比を非円形度 $r = a_l/a_s$ とし，慣性主軸方向の単位ベクトルを (u,v) として，$\{u(1-r), v(1-r)\}$ で手の方向を表すことにしている．

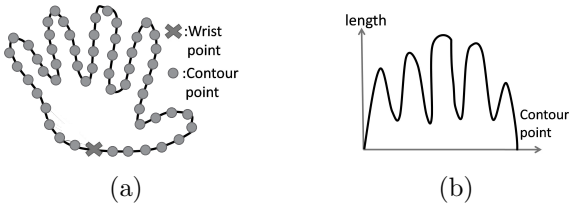


図 7　突起数
Fig. 7　Number of protrusion.

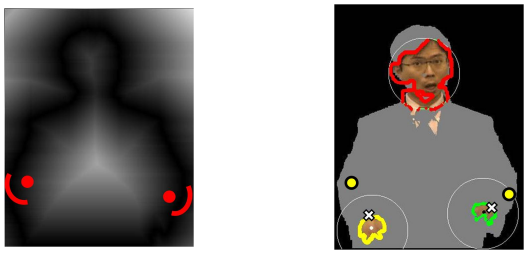


図 8　肘点, 手首点の検出
Fig. 8　Detection of elbow points and wrist points.

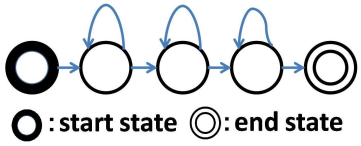


図 9　left-to-right HMM
Fig. 9　A left-to-right HMM.

4.　手話単語の学習

これまでに述べた手法で得た特徴を用いて，単語毎に HMM モデルを作成し，学習，認識を行う．

4.1　HMM モデルによる学習と認識

本手法で利用する HMM は図 9 のような Left-to-Right モデルである．入力から得た特徴時系列データから各単語モデルの尤度を Viterbi アルゴリズムにより計算し，最大尤度を持つ単語を認識結果とする．ここでは特徴の確率分布として多次元ガウス分布を仮定し，各単語モデルの学習には Baum-Welch アルゴリズムを用いる．

単語モデルの学習には，モデルの状態数を仮定し，さらにモデルパラメータの初期値を与える必要がある．モデルパラメータは，各状態の特徴の確率分布と状態

遷移確率からなり，それらを Baum-Welch アルゴリズムで学習するには，あらかじめ状態数と状態間の遷移構造を定めておく必要がある．

従来は単語ごとに手動で状態分割するか，全単語で共通の状態数を与え，パラメータの初期値を設定している．しかし，実用を考えた際に手話単語数は膨大なため，手動で全単語の状態分割を行うのは困難である．また，全単語に共通の状態数を与える手法では，単語毎の複雑さの違いを考慮できず，モデルの記述力不足や過学習による汎化性能の低下が起こる．そこで，自動で特徴時系列データの適切な状態分割を行い，初期値を設定する手法を提案する．

なお，ここでの状態分割は，HMM 学習の初期値を与えるものであり，学習サンプルに対して必ずしも最大尤度を与えるものではない．学習サンプルに対する尤度だけでは状態遷移構造を決定できないため，本論文では構造を Left-to-Right に制限し，状態分割によって状態数とモデルパラメータの初期値を決定した後，Baum-Welch アルゴリズムによって学習サンプルに対する尤度を最大化する方法を採る．状態分割によって単語動作を構成する単純な動きの数が得られればそのそれぞれに対応した状態を持つ初期モデルを採用することでひとつの状態がひとつの動きに対応した初期 HMM を生成できる．

4.2 画像系列の状態分割

手話の特徴として，手の速度が大きい時は手の動きが重要であり，速度が小さい時は手の形状が重要である [15]．そこで，手の速度によって運動区間と静止区間に状態分割する．運動区間の中でも，手の速度が大きく変化する場合は運動の性質が変わることが多い．手の速度が大きく変わるところでは運動区間を分割する．さらに，運動区間を手の運動方向の変化により分割する．顔の近くで行われる手話の場合，手の動きが小さく，単純な速度による分割では状態分割できないことがある．顔と手の距離が小さく，顔からの方位角が大きく変化する区間を運動区間ととらえ，分割を行う．

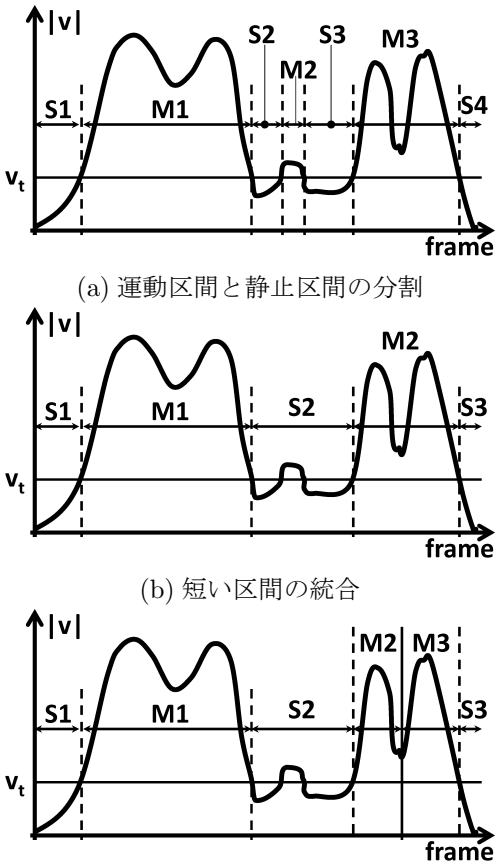
まず，片手だけの場合の状態分割について述べる．

1. 速度による分割

(a) 運動区間と静止区間の分割

XY 座標系における手の重心の速度ベクトルを $\vec{\mathbf{v}} = (v_x, v_y)$ とする．$|\vec{\mathbf{v}}|$ が閾値 v_t 以上の区間を運動区間 (M)，閾値未満の区間を静止区間 (S) として分割する (図 10-(a))．閾値 v_t は，手話の速度に依存するが，ここでは前述の手の長軸の長さ d_p を基準として，$v_t = 2d_p$[pixel/sec] とした．図中の速度のグラフの下にある S1 や M2 が分割結果である．

(b) 短い区間の統合



(a) 運動区間と静止区間の分割

(b) 短い区間の統合

(c) 運動区間の分割

図 10 手の速度による画像系列の分割

Fig. 10 Segmentation by the hand motion speed.

短い区間（0.1sec 以下）で，速度の最大値 $|\vec{\mathbf{v}}_M|$ が $|\vec{\mathbf{v}}_M| < v_t + \Delta v$ を満たすような運動区間は前後の静止区間と統合する ($\Delta v = 0.6v_t$) (図 10-(b))．

(c) 運動区間の分割

運動区間内で $|\vec{\mathbf{v}}|$ が極小値をとり，両隣の極大値との差がどちらも閾値 d_m^v 以上となる場合は，極小値をとる点で区間を分割する ($d_m^v = 0.1d_p$[pixel/sec])(図 10-(c))．

2. 運動方向による分割

$|\vec{\mathbf{v}}|$ が極小値をとり，その前後 0.1[sec] 以内に手の運動方向が閾値 α_t 以上変化するならばそこで区間を分割する ($\alpha_t = 2$[rad])(図 11)．

3. 顔からの方位角による分割

手話では目，鼻，口など顔の近くの位置が重要である．そこで，顔と手のユークリッド距離 r' が d_p 以下で，顔からの方位角が $\pi/8$[rad] 以上変化する場合は，その中間点で分割する．

なお，以上で用いた閾値は，厳密には話者に依存するが，HMM の学習の初期値として用いる限り，認識に大きな影響を与えない．

次に，両手の分割手法について述べる．まず，片手

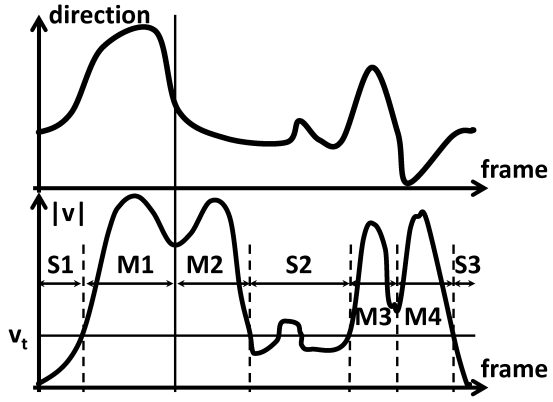


図 11　向きによる状態分割
Fig. 11　Division by the direction.

ずつ上述の 1 から 3 の処理を行い，特徴時系列データを分割する．両手の手話はそれぞれの手を同時に動かすことが多いが，タイミングに少しずれが起こる場合がある．そこで，左右の状態分割点のずれが小さい場合（0.1[sec] 以内）は同時とみなし，状態の過分割を防ぐ．実際の手話動画の分割例を図 12 に示す．図中の上のグラフは手の動きの向きを，下のグラフは手の速さを表している．下の (1) から (5) の画像は状態分割点での原画像である．

得られた状態分割結果を用いてモデルパラメータの初期値を設定する．遷移確率は，同じ状態に遷移する確率を 0.6，次状態に遷移する確率を 0.4 とする．各状態の特徴の確率分布は，その状態に属するフレームの特徴の平均及び分散を初期値とする．但し，フレーム数の少ない静止状態で分散が極端に小さくなるのを防ぐために最小分散を導入し，実際の特徴から求めた分散がこれを下回る場合には分散の値を最小分散に置き換える．

5.　実験

提案手法の有効性を確かめるため，2 つの実験を行った．まず，提案した特徴表現の有効性を確かめるため，XY 座標系と認識性能の比較実験を行った．次に，提案した状態分割法の性能を確かめるため，全単語に共通の状態数を与えて学習を行った場合と認識性能の比較実験を行った．本章ではそれぞれの実験の内容と結果・考察について述べる．

5.1　実験対象

実験対象には買い物の場面を想定した約 50 語に手話技能検定 6,7 級から選んだ平易な語を加えた，表 1 に示す合計 92 単語，480 動画 (以後サンプルと呼ぶ) を用いた．これらの動画は 640×480[pixel], 30[frame/sec] で 3 秒から 5 秒程度の長さで，正面から話者が撮影されたものである．実験対象に現れる手形状の例を図 13 に示す．動作はある位置で静止するような語が 13 語，

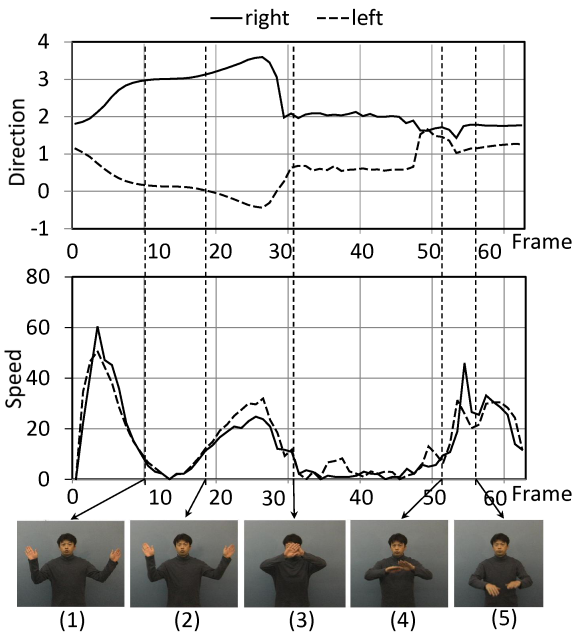


図 12　状態分割の例
Fig. 12　Example of state division.

図 13　実験対象の手形状例
Fig. 13　Example of hand shapes.

単純な 1 つの動作を含む語が 54 語，2 つ以上の動作を含む語が 27 語ある（同一の単語が複数の分類に属する場合がある）．また，480 サンプル中 238 サンプルに顔と手もしくは手同士の隠蔽がある．本論文では手話による聴覚障害者との会話を経験している者を熟練者，その経験のない者を非熟練者と呼ぶこととし，熟練者 7 名と非熟練者 3 名による発話を用いて実験を行った．単語によって熟練者と非熟練者の割合は異なるが 6 語 (北，東，西，あなた，私，来週) は非熟練者から得たサンプルからのみなり，残りの単語は熟練者が半分以上の割合を占めている．

これらの語に対し特徴抽出を行ったところ，目視で正しく追跡が出来たサンプルは 446 サンプルであり，追跡に失敗したサンプルは全て隠蔽を含むものであった．隠蔽を含むサンプルに限定した場合の成功率は 85%以上であり，提案した手領域追跡法は隠蔽にも有効に対処できるものと言える．今回は追跡に失敗したサンプルを除き，サンプル数が 4 例以上あった表 2 に示す 55 語を 5.2 節，5.3 節の認識実験に用いた．認識

表 1 実験対象単語
Table 1 Set of words used in experiment.

明日	明後日	歩く	上	嬉しい
覚える	重い	きれい	夜	狭い
晴れ	昼	広い	毎日	待つ
金曜日	スカート	ズボン	シャツ	靴
ネクタイ	セーター	眼鏡	色	赤
青	黒	白	絹	皮
茶色	緑	黄	腕	指
薄い	厚い	小さい	大きい	長い
短い	円	春物	夏物	冬物
涼しい	暖かい	暑い	寒い	肩
胸	頭	安い	高い	流行
背が高い	背が低い	好き	嫌い	値上げ
ありますか	東	西	北	cm
どこですか	終わる	右	左	来週
していいですか	似合う	教える	飲む	かばん
～を下さい	値下げ	もっと	顔	雨
ありがとう	下	一昨日	昨日	日曜日
月曜日	火曜日	水曜日	木曜日	土曜日
私	あなた			

表 2 認識実験に用いた単語
Table 2 Set of words used in recognition experiment.

明日	していいですか	似合う	上	来週
教える	重い	きれい	右	狭い
晴れ	飲む	広い	毎日	待つ
金曜日	スカート	ズボン	シャツ	靴
ネクタイ	セーター	眼鏡	色	赤
青	黒	白	絹	皮
かばん	厚い	小さい	大きい	長い
短い	cm	円	夏物	冬物
涼しい	暖かい	暑い	寒い	肩
左	頭	安い	終わる	流行
東	背が低い	好き	私	あなた

実験に用いたサンプルには指の一部が欠けたり，隣接する別の手・顔領域の一部を含んでしまったりするものも含まれているが，このような失敗は数フレームであり，HMM を認識に用いることで誤差を吸収できると期待されるので実験に用いて認識性能の評価を行っている．

5.2 特徴の違いによる認識性能の比較実験

3 章で提案した対数極座標表現による位置の特徴の有効性を調べるため，提案手法による手の重心位置の特徴 (r,θ) 及び $(\Delta r,\Delta\theta)$ の代わりに顔を基準とした XY 座標系で表現された手の重心位置の特徴 (X,Y)，及び $(\Delta X,\Delta Y)$ を特徴としたものと比較した．その他の特徴 (x,y),N_r,N_l，$\{u(1-r),v(1-r)\}_r$,$\{u(1-r),v(1-r)\}_l$ はどちらも使用している．HMM 生成時の最小分散は実験的に 1.0×10^{-6} とした．

各単語につき 4 つの時系列画像データを準備する．各単語から時系列データを 1 つずつ集めたデータセットをグループとし，時系列データが重複しないグループを 4 つ作成する．実験では認識対象として 1 グループ，学習用として 3 グループを割り当て，クロスバリデーションを行った．ただし，各実験では 3 つの学習用グループのうち 1 つに対して提案手法による状態分割を用いて初期モデルを作成し，学習用の 3 グループの時系列データで各単語の HMM を作成した．認識対象グループと初期モデル生成用グループの割り当て方は $4\times3=12$ 通りあるので，その全てについて実験を行った．

表 3 特徴による認識率の比較
Table 3 Comparison of recognition rate between different features.

特徴	XY 座標系 $(X,Y,\Delta X,\Delta Y)$	対数極座標系 $(r,\theta,\Delta r,\Delta\theta)$
認識率	66.8% (441 例/660 例)	75.9% (501 例/660 例)

尤度一位の単語 HMM が認識対象の正解単語と合致すれば認識に成功として，全てのグループ割り当て (12 通り) と単語 (55 単語) についての平均認識率を求めた結果を表 3 に示す．XY 座標系を用いる手法に比べ，提案手法は誤認識が約 10%改善した．改善した単語の例を図 14 に示す．誤認識が改善された 29 語のうち顔との隠蔽を含む単語が 6 語 (図 14(a))，顔付近で動作が行われた単語が 10 語 (図 14(b))，顔から離れた位置で行われた単語が 11 語 (図 14(c)) あった．このことから，顔付近で行われる単語だけでなく，顔から離れた位置で行われる単語にも提案手法が有効であるといえる．

また,「明日」と「来週」,「短い」と「小さい」,「長い」と「大きい」,「重い」と「終わる」，これらの語は動きが似ており形状特徴無しには区別が難しい語であるが，これらの語は 79.6%認識に成功した．このことより，提案法による形状特徴は手話認識に有効な特徴であると言える．

5.3 状態分割法による認識性能の比較実験

次に，提案した状態分割法の有効性を確かめるため，以下の条件で状態分割を行い，作成したモデルの認識性能の比較を行った．

- 提案法による状態分割 (A)
- 全単語で同一の状態数を与え各サンプルの時系列データを等分割 (B)

提案法ではデータセットに対し平均 4.7 状態に分割した．同一単語でも推定状態数に差があるものが多かったが，動作の前後の静止状態の有無がほとんどであった．(B) については，状態数を 3 から 13 まで変化させてそれぞれで認識実験を行った．なお，認識に用いる特徴は提案法のものを用いた．

実験結果を図 15 に示す．図中の正解率は尤度 1 位の単語モデルが認識対象の単語と合致した割合を示す．

(a) 顔と手の隠蔽がある単語

(b) 顔付近で動作が行われた単語

(c) 顔から離れた位置で行われた単語

図 14　誤認識が改善した例
Fig. 14　An example of word which recognition rate has been improved.

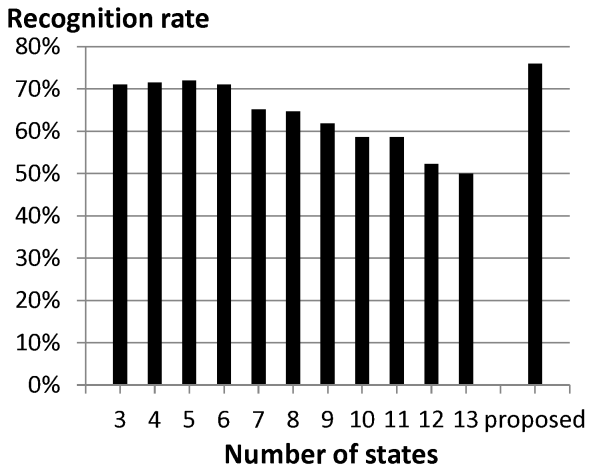


図 15　状態数固定と提案手法の認識率の比較
Fig. 15　Comparison of recognition rate between fixed state number and the proposed method.

提案法による認識性能（75.9%）と，(B) の内最高の認識性能 (状態数 5,71.9%) がほぼ同等の結果となった．特定サンプルの状態分割結果を採用する提案法では適切な状態数とならない恐れはあるが，この実験結果を見ると状態数を固定するより高い認識性能が得られることが期待できる．また (B) の手法は認識対象の語彙に対し，どの状態数が最適か実験的に試行錯誤する必要があるが，提案法ではその必要がない．語彙が多い時に提案法は特に有用であると期待できる．

5.4　提案法の誤認識の原因

提案手法の誤認識の原因は大きく分けて 3 つ考えられる．

図 16　手の位置・動きが大きく異なる同一単語
Fig. 16　Big difference in start position and motion among the same words.

1 つ目は特徴抽出の失敗である．

1-a) 位置に関する特徴抽出の失敗

位置の特徴がうまく抽出できない原因はテンプレートマッチングによる誤差である．テンプレートと見えが少し変わる場合，ずれが生じる場合があった．

1-b) 形状に関する特徴抽出の失敗

形状に関する特徴抽出失敗は指を前後方向に伸ばしている時に起こる．本手法では手領域の輪郭から突起数を計算しているため，突起数が計算できない場合がある．

2 つ目はサンプルのばらつきである．今回の実験では学習データが一単語につき三つのため，図 16 のように動きに大きな個人差が見られる場合に HMM が持つ統計的な性質を利用できず失敗したと考えられるものがあった．これらの問題はサンプル数を増やせば対応できると考えられる．

3 つ目は提案法で用いた HMM モデルでは表現が難しい単語である．提案法では「暖かい」のような動作の繰り返しがある単語に対し，1 回の繰り返し動作を複数の状態に分割する場合がある．サンプルによって繰り返し数が異なる場合，Left-to-Right HMM モデルでは表現が難しい．より複雑な状態構造を持つ単語モデルの構築は今後の課題である．

6.　むすびに

本論文では，画像系列から得られる特徴を用いて手話単語の学習・認識を行う手法を提案した．特徴抽出では，手同士や顔と手の重複が起こってもテンプレートマッチングや明度相関を利用し，位置を検出し追跡を行った．手・顔領域の追跡結果より，見えを考慮した形状の特徴を抽出した．位置に関する特徴は顔を基準とした極座標系で，距離の対数をとることで手の位置の細かな違いによる誤認識の改善ができた．また，手の運動から単語モデルの状態数の自動推定を行う手法を提案し，従来法の実験的に最適な状態数を選択する場合と同等の性能を出せることを実験的に示した．

今後の課題は，特徴抽出実験結果で述べたように特徴抽出の不十分なところへの対処，および未知の手話に対する認識性能を高めるための HMM の状態構造の最適化である．

謝辞

この論文を執筆するにあたり，手話の動画を提供してくださった中京大学の神田和幸教授，及び手話技能検定協会の方々に深く御礼申し上げます．

参考文献

[1] 中山武明, 福村直博, 宇野洋二: ヒトの身振り認識過程を考慮した前腕躍度による手話単語認識システムの検討; 信学技報. NC2004-167, pp.179-184 (2005)
[2] 高岡哲也, 福村直博: 運動の滑らかさと単語運動時間変動を考慮した手話動作の解析; 信学技報, NC2006-203, pp. 91-96 (2007)
[3] 田中了, 石川眞澄: 複数の自己組織化マップに基づく手話動作認識; 信学技報,NC2002-216,pp.77-82 (2003)
[4] Yoshino, L., Kawashima, T., Aoki, Y.: Recognition of Japanese sign language from image sequence using color combination; Proc. 3rd Int. Conf. Image Processing, Vol.3, pp.511-514 (1996)
[5] 吉冨康成, 永山しづえ, 杉山雅祥: 温度画像処理による手軌跡の抽出と手話認識; 信学技報,SP2002-112,pp.21-26 (2002)
[6] 岡澤裕二, 堀内靖雄, 市川熹: オプティカルフローによる手話の大局的動作の認識について; 信学技報,PRMU2002-77,pp.39-44 (2002)
[7] 星野聖, 小渡悟, 神里志穂子, 新垣武士: 手話認識のための 3 次元手指運動推定; 信学技報,NLP2000-170,pp.43-49 (2001)
[8] 李昌宏, 中薗薫, 長嶋祐二, 張鴻徳: 動きベクトルを用いた手話単語分類; 信学技報, WIT2003-62, pp.65-70 (2004)
[9] 木場洋介, 石川眞澄: 動画像からの移動物体抽出と手話動作認識への応用; 信学技報, NC2002-170,pp.221-226 (2003)
[10] 今川和幸, 谷口倫一郎, 有田大作, 松尾英明, 呂山, 猪木誠二: カメラを用いた手話認識における見えの違いを考慮した手話の局所特徴認識; 映像情報学会誌,Vol.54,No.6,pp.848-857 (2000)
[11] 山本貢嗣, 野村健, 南角吉彦, 後藤富朗, 北村正: HMM に基づく日本手話認識のための特徴の統合に関する検討; 電子情報通信学会総合大会講演論文集, A-4-28, p.97 (2005)
[12] 柳哲, 柳生雄午, 徳田恵一, 北村正: 手の動作と形状を用いた HMM 手話認識; 電子情報通信学会総合大会講演論文集,D-12-119,p.285 (2004)
[13] Starner, T., Pentland A., Weaver, J.: Real-time American Sign Language recognition using desk and wearable computer based video; IEEE Trans. PAMI, Vol.20, No.12, pp.1371-1375 (1998)
[14] Ong, S.C.W., Ranganath, S.: Automatic Sign language Analysis: A Survey and the Future beyond Lexical Meaning; IEEE Trans. PAMI, Vol.27, No.6, pp.873-891 (2005)
[15] 佐川浩彦: 手話認識における手動作セグメンテーション方式; ヒューマンインタフェースシンポジウム'99 論文集,pp.749-754 (1999)
[16] 金山和功, 白井良明, 島田伸敬: HMM を用いた手話単語の認識; 信学技報, PRMU2004-16, pp.21-28 (2004)
[17] 長嶋祐二, 神田和幸: 手話のコンピュータ処理; 電子情報通信学会誌, Vol.84, No.5, pp.320-324 (2001)

（2012 年 3 月 21 日受付，10 月 19 日再受付）

著者紹介

松尾　直志

2001 年 京都工芸繊維大学 電子情報工学科卒．2006 年 同大大学院 工芸科学研究科博士課程了．2006 年立命館大学総合理工学研究機構研究員．2012 年より同情報理工学部知能情報学科助手．画像認識の研究に従事．電子情報通信学会会員．

山田　　寛　（学生会員）

2007 年 立命館大学理工学部 情報学科卒．2009 年 同大大学院 理工学研究科博士前期課程了．2009 年より立命館大学大学院 理工学研究科博士後期課程．手話認識の研究に従事．電子情報通信学会会員．

白井　良明　（正会員）

1964 年名古屋大学工学部機械工学科卒業．1969 年東京大学大学院工学系博士課程修了．工学博士．同年電子技術総合研究所入所．1988 年大阪大学工学部教授．2005 年立命館大学理工学部教授．2012 年より同総合科学技術研究機構客員教授．コンピュータビジョン，知能ロボットなどの研究に従事．人工知能学会，情報処理学会，日本機械学会，電子情報通信学会，IEEE など各会員．

島田　伸敬

1992 年大阪大学工学部電子制御機械工学科卒．1997 年大阪大学大学院博士後期課程了．博士 (工学)．同年大阪大学大学院工学研究科助手．2003 年同助教授を経て，2004 年立命館大学情報理工学部准教授．2012 年より同学部教授．2007 年より 1 年間米カーネギーメロン大学ロボティクス研究所客員准教授．コンピュータビジョン，ジェスチャインターフェース，対話ロボットの研究に従事．電子情報通信学会，情報処理学会，人工知能学会，IEEE 各会員．