コンピュータビジョンと
イメージメディア 127－18
（２００１．５．１０）

自己組織マップと隠れマルコフモデルを用いたジェスチャ認識の研究 —リップリーディング・システムの試作—

井内浩貴　前田佐嘉志　鶴田直之
福岡大学大学院　工学研究科

ジェスチャ認識において、人のジェスチャを離散的な状態の遷移に変換する際に、時々遷移の途中で誤認識を起こす場合がある。そこで本論文では、そのような誤認識を減らすために特徴のパターンを確率的に捉えた隠れマルコフモデル (HMM) を用いたジェスチャ認識について論じる。また、離散的な画像の生成と、HMMにおける状態遷移系列の生成には、ベクトル量子化を利用した自己組織マップ (SOM) を用いて、SOMとHMMを併用することにより認識率の向上を目指す。SOMとHMMを用いたジェスチャ認識実験として、口の動きをカメラで取り込み、コンピュータに認識させるリップリーディング・システムを試み、その認識率について検討する。

Gesture Recognition using Self-Organizing Maps and Hidden Markov Model —Trying of Lip Reading System—

Hirotaka Iuchi, Sakashi Maeda, Naoyuki Tsuruta
Graduate School of Engineering, Fukuoka University

Gesture recognition is an appealing tool for natural interface with computers especially for physically impaired persons. In this paper, it is proposed to use Self-organizing maps (SOM) as an image recognition system for Hidden Markov Model (HMM) base gesture recognition, since the SOM allows alleviating many difficulties associated with gesture recognition. By vector quantization using the SOM, discrete states, which are useful for stochastic analysis of the HMM, are generated. In addition, to reduce the recognition time to the range of normal video camera rates can be achieved. An experimental result to read lip motion using the SOM is presented.

1 序論

現代社会において、お年寄りや子供を含めた多くの人とコンピュータは深く関わりを持つようになってきている。その反面、入力インターフェイスであるキーボード、マウスなどの扱いに悩まされ、利用しきれない人が多くいることも事実である。

そこで、本研究ではコンピュータによる人のジェスチャ認識について検討する。コンピュータが人間のジェスチャを認識できるようになれば、コンピュータの難しい操作をしなくとも人間の意志伝達等が可能となる。つまり、キーボードやマウスなどを扱わない、いわば非接触型インターフェイスが実現し、お年寄りや子供でも気軽にコンピュータと関わることができるようになるだろう。

手や口の動きなどを認識する場合、人のジェスチャを離散的な状態の遷移に変換してジェスチャ認識を行う。そこで必要となってくるのは、離散的な状態を予め作っておくことである。その離散的な状態を作ったり、人間のジェスチャを離散的な状態の遷移に変換してジェスチャ認識を行ったりする手法として、ベクトル量子化を用いた手法である自己組織マップ (SOM) がある。その SOM に用いる画像を取り込む際には、口の位置や姿勢を注意深く一致させなければならないという問題がある。しかし、この問題は口の位置ずれなどの問題に対応したハイパーコラムモデル (HCM) を用いることにより解消することができる [4][5]。本研究の実験では SOM を用いているが、今後 SOM の代わりに HCM を用いると認識率が向上すると考えられる。

実際に入力データが与えられた時、もしその入力データの中に極稀に起こる誤りのデータが含まれていたら、SOM による認識だけではその誤りのデータも他のデータとともに認識してしまい、結果的には誤った認識をしてしまう。そこで、そのような誤認識を減らし、入力データをより正しく認識する手法として、HMM(Hidden-Markov-Model) がある。

本研究では、人間のジェスチャ認識の一つとして、リップリーディングを試みる。人間の口の動きをコンピュータが認識することができれば、コンピュータを通じて人の意志を伝達することができる。そこで、ジェスチャのパターン認識に使われる手法の一つで音声認識で実績にある HMM と SOM を用いてリップリーディングを行い、その認識率について検討する。

2 ジェスチャ認識

2.1 状態系列のマッチング

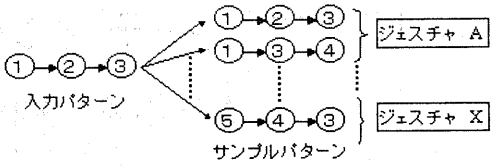

図 1: ジェスチャ認識

ジェスチャはいくつかの状態変化の連続であり、ジェスチャ認識はジェスチャを離散的な状態の遷移に変換して行う。図 1 のように、入力パターンの状態遷移とサンプルパターンの状態遷移を比較し優勝者 (最も類似度が高い状態遷移) を選出し、それが属するジェスチャを認識結果とする。

2.2 自己組織マップ (SOM)

自己組織マップ (SOM) の動作原理を示す。データ空間に M 個の標本 $\{i|i=1,2,..,M\}$ をランダムに配置する。その標本画像 i が持つ標本ベクトル (標本画像) を $\mathbf{m}_i$ と表す。自己組織マップ (SOM) の標本画像は 1 次元配列に順序良く環状に配置されていて、隣り合う画像同士は類似している。標本の位置ベクトルを $\mathbf{R}_i$ で表す。X 個の入力画像 $(x=1,2,..,X)$ の入力ベクトルを $\mathbf{x}_j(j=1,2,..,X)$ とすると、学習はまず、時刻 $t=0$ において、入力画像に対し全ての標本画像との類似度をユークリッド距離 D を用いて計算する。

$$D = ||\mathbf{x}_j(t)-\mathbf{m}_i(t)|| \quad (1)$$

$$= \left((\mathbf{x}_j-\mathbf{m}_i)^2\right)^{1/2} \quad (2)$$

$$1 \le j \le X, \quad 1 \le i \le M.$$

D が最小となる標本画像 (各入力画像に対して最も類似度が高い標本画像) を勝者の標本画像 c とするとその勝者の標本画像 c の近傍の標本画像に対しても以下の式に基づいて標本ベクトルを更新する。

$$\mathbf{m}_c(t+1) = \mathbf{m}_c(t) + h_{c.i}[\mathbf{x}_j(t) - \mathbf{m}_c(t)], \quad (3)$$

$$h_{c.i} = \alpha(t)\beta(t), \quad (4)$$

$$\beta(t) = \exp\left(-\frac{||\mathbf{R}_c - \mathbf{R}_i||^2}{2\sigma^2(t)}\right). \quad (5)$$

ただし、$\alpha(t)$ と $\sigma(t)$ は以下の通りである。

$$\alpha(t) = 1.0 - \frac{t}{\mathbf{t}_{max}}, \quad (6)$$

$$\sigma(t) = \frac{M}{2.0} + \left(\frac{\frac{1.0-M}{2.0}}{\mathbf{t}_{max} \times t}\right). \quad (7)$$

$\alpha(t)$ は学習速度であり、どの程度入力画像に近づけるかを調整するパラメータである。

以上の処理について時間を、1 つずつ増やしながら時刻 $t = \mathbf{t}_{max}$ まで繰り返す。

SOM の認識においては、ある入力画像が与えられたとき、SOM 学習によって得られた標本画像と入力画像の類似度をユークリッド距離を用いて計算し、最も類似度が高い標本画像を勝者として出力する。

2.2.1 自己組織マップも用いる理由

- 自己組織マップ (SOM) で生成された標本画像は 1 次元上で環状に配置される。そして、隣あう画像同士は類似している。この性質を利用した高速化により、実時間処理が可能である [6]。

- 自己組織マップ (SOM) の学習の繰り返し処理中において、入力画像に最も類似度が高かった標本画像 (勝者の標本画像) だけでなく 1 次元上に配置された近傍の標本画像も同じような学習を行うため、自己組織マップ (SOM) の学習では非参照標本画像が生成されにくい。これにより、最適なベクトル量子化をよく近似することができる。

以上の点が、離散的な状態の生成と HMM における状態遷移系列の生成のために自己組織マップを用いる理由として挙げられる。

3 隠れマルコフモデル (HMM)

3.1 HMM とは

動画像を用いた認識手法のうち、時系列データを扱うものとして隠れマルコフモデル (HMM) がある。HMM は記号出力や状態遷移を確率的に考えたモデルである。また、大量の時系列データを統計的に扱えるため学習が可能であり、不特定者認識に対して有効であることが示されている。この HMM は音声認識の分野で成果を上げているが、近年では画像認識の分野においても広く用いられている。

3.2 HMM パラメータについて

HMM には 5 つのパラメータがあり、以下に示す値で特徴づけられている。

- モデルの状態数：N
- 出力記号数：M
- 状態遷移確率分布:$A = \{a_{ij}\}$ ， $a_{ij} = P(q_{t+1} = j|q_t = i)$ ， $1 \leq i$ ， $j \leq N$

  a_{ij} は状態 q_i から状態 q_j への遷移確率であり、$\sum_j a_{ij} = 1$ を満たす。
- 記号出力確率分布：$B = \{b_i(v_k)\}$ ， $b_i(v_k) = P(o_t = v_k|q_t = i)$ ， $1 \leq k \leq N$

  $b_i(v_k)$ は状態 q_i で記号 v_k を出力する確率であり、$\sum_t b_i(v_k) = 1$ を満たす。
- 初期状態確率分布：$\pi = \{\pi_i\}$ ， $\pi_i = P(q_1 = j)$ ， $1 \leq i \leq N$

  π_i は状態 q_i が初期状態である確率である。

以下ではモデルのパラメータを $\lambda = (N, M, A, B, \pi)$ と表記する。

3.3 HMM の基本原理

例として図 2 のような N 個の状態 1,2,....N を各時刻においてとるシステムを考える。このシステムでは、以下のような手順で出力記号系列 $\mathbf{O} = (o_1, o_2, o_3, ..., o_T)$ を出力する。

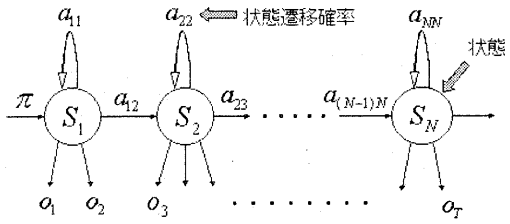

図 2: 隠れマルコフモデル (left-to-right モデル)

1. t=1 とし、初期状態確率分布 π に従い、初期状態 q_1 を選ぶ。また、状態 q_1 における、記号出力確率分布 $b_{q_1}(k)$ に従い、出力記号 v_k を o_1 として選ぶ。
2. 状態 q_t における状態遷移確率分布 $a_{q_t q_{t+1}}$ に従い、新しい状態 q_{t+1} へと遷移する。
3. 状態 q_{t+1} における、記号出力確率分布 $b_{q_{t+1}}(k)$ に従い、$o_{t+1} = v_k$ を選ぶ。
4. t=t+1 とし、$t < T$ ならば、2 に戻り、そうでなければ手続きを終える。

このシステムでは、ある状態からの出力記号だけが観測され、状態遷移は直接には観測されないため、「隠れマルコフモデル」と呼ばれている。

3.4 HMM によるジェスチャ認識

HMM によるジェスチャ認識は、サンプルデータから予め HMM パラメータを学習させておく。観測された遷移系列 $\mathbf{O} = (o_1, o_2, ..., o_T)$ が入力したジェスチャである確率 $P(\mathbf{O}|\lambda)$ を HMM パラメータをもとに算出し、その確率が最大となる HMM を認識結果とする。

観測系列 $\mathbf{O} = (o_1, o_2, ..., o_T)$ に対する最適な状態遷移系列とは $P(\mathbf{O}, \mathbf{q}|\lambda)$ を最大化するような系列 $\mathbf{q} = (q_1, q_2, ..., q_T)$ である。この最良な唯一の状態系列をみつける手法として Viterbi アルゴリズムを用いる。

3.5 HMM の特徴

HMM はパターン変動を確率モデルで捉え、統計的処理ができるので不特定認識に対して有理である点などが利点として挙げられるが、欠点としては、モデルの設計法が確立していない点や、HMM パラメータ推定に多量の学習サンプルを要する点が挙げられる。

4 認識アルゴリズム

4.1 前提条件

リップリーディングを試みる際の前提条件として、基本となる状態を決めておく必要がある。そこで本研究では、基本となる状態として、母音『あ』、『い』、『う』、『え』、『お』と『ん』の 6 つの状態を HMM における状態数とする。また、サンプルとして、『頭』、『胸』、『お腹』、『背中』、『手足』、『いたい』、『くるしい』、『おもい』、『すいた』、『しびれる』の 10 単語の動画像を用いて、『頭いたい』、『頭おもい』、『胸いたい』、『胸くるしい』、『お腹いたい』、『お腹すいた』、『背中いたい』、『手足いたい』、『手足しびれる』の 9 つの口の動作の認識を試みる。そして認識の際には、『頭』→『あ』『あ』『あ』、『いたい』→『い』『あ』『い』というように各単語は母音の遷移として考える。

取り込む画像としては、口周辺の画像を取り込む。そして、実際に SOM で学習、認識する際には、口の周辺部分のみの類似度計算を行い、その他の部分 (背景など) には依存しないようにする。

4.2 リップリーディングアルゴリズム

実際にカメラで取り込んだ動画像を扱う際には、動画像を離散的な画像の遷移に変換して、ジェスチャ認識を行う。その変換の手法として SOM を用いる。取り込んだ動画像と予め作っておいた標本画像の類似度計算を各フレームごとに SOM を用いて行い、それぞれ勝者の標本を選ぶ。その標本画像の遷移系列を入力データの観測系列とする (図 3)。

観測系列が得られたら、その観測系列により HMM パラメータを推定する。そして、得られた観測系列がどういう状態遷移系列から生成されたかを確率的に求める手法として HMM を用いる。実際、得られた観測系列を生成する状態系列は複数あり、観測系列を生成する確率が最も高い状態遷移系列を求める方法として、Viterbi アルゴリズムを用いる。そ

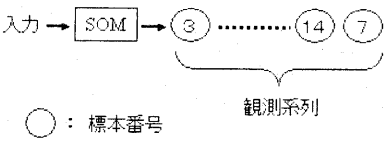

図 3: 入力データ観測

して、最も確率の高い状態遷移系列を最終的な状態遷移とする (図 4)。

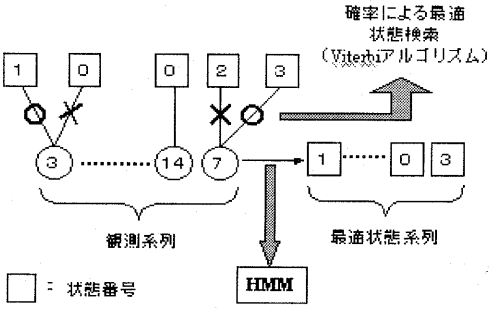

図 4: 最適状態遷移検索

リップリーディングアルゴリズムを以下にまとめておく。

1. L 枚の口のサンプル画像から SOM 学習により、M 枚の標本画像を作る。
2. いくつかの口の動画像をサンプル入力データとして取り込み、SOM 認識 (入力画像と標本画像の類似度計算) によりサンプル入力データを離散的な画像の遷移 (標本画像の遷移) として観測する。
3. サンプルデータから得られたいくつかの観測系列から、HMM パラメータを求める。
4. 実際に口の動画像を入力し、Viterbi アルゴリズムにより最適な状態遷移系列を求める。

5 ジェスチャ認識実験

5.1 実験の目的

実際に口の動画像を用いてジェスチャ認識を行い、SOM 認識のみの場合と HMM を用いた場合との認識率の比較し、HMM の有効性を検討する。また、入力データに HMM 学習に用いた画像と HMM 学習未使用画像を用いた場合の認識率を比較し、HMM が全体的にどのくらい有効かどうかを検討する。

5.2 実験の条件

本研究では、SOM 学習に 1172 枚の口の画像を用いて、31 枚の標本画像を作る (図 5)。用意した 10 単語の動画像を用いて、9 種類のジェスチャ 7 人分を認識実験に使用する。また今回の実験は、白黒 256 階調、160 × 120 ピクセル、30 フレーム/秒の画像を用いる。

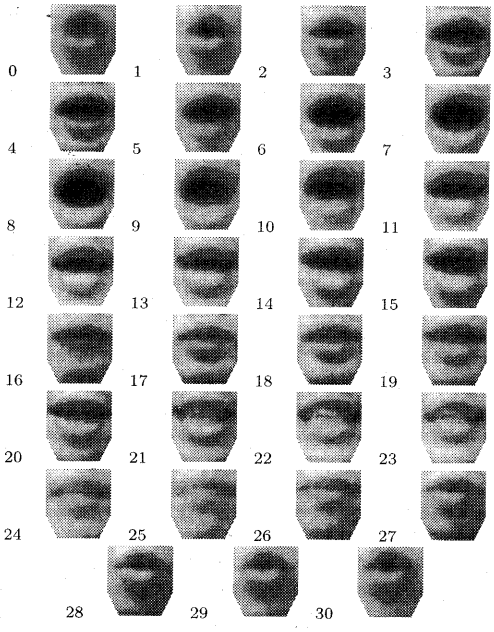

図 5: 標本画像

認識実験 1 として、7 人分× 9 = 63 個のジェスチャを用いて、SOM による認識だけの場合の認識実験を行なう。それから認識実験 2 として、7 名の

SOM(学習)	サンプル画像	1172 枚
	繰り返し回数	1500 回
	標本枚数	31 枚
HMM	単語数	10
	動作内容	ジェスチャ 9 種類
	人数 (口の種類)	7
	モデル (パラメータ推定用)	9 モデル × 4 人分 = 36 モデル
	モデル (実験用)	36 モデル + (9 モデル × 3 人分) = 63 モデル
	状態数	6 (『あ』、『い』、『う』、『え』、『お』、『ん』)
	標本枚数	31 枚
共通データ	画像サイズ	横 160 × 縦 120 pixel
	階調数	gray scale , 256 階調
	フレーム周期	1/30 sec
	1 動作フレーム数	90 ～ 140

表 1：実験条件

被験者のうち 4 人分のデータをモデルの学習 (パラメータ推定) の際の学習データに使用し、そのサンプルデータを用いて認識実験して得られた認識率を 1 つの実験結果とする。そして、学習データに使用しなかった残り 3 人分のデータを用いて、同様の認識実験を行ない、その認識率をもう 1 つの実験結果とする。

実験条件についてまとめたものを表 1 に示す。

5.3 実験結果

5.3.1 認識実験 1

『あ』、『い』、『う』、『え』、『お』、『ん』の 6 つの状態の適当な画像を用意し、その画像と作った標本画像をそれぞれ類似度計算して標本画像を分類すると、6 番から 11 番は『あ』、21 番から 23 番は『い』、24 番と 28 番から 30 番は『う』、3 番から 5 番と 12 番から 14 番と 20 番は『え』、16 番から 19 番と 25 番から 27 番は『ん』のように分類された。

SOM のみの認識の場合は、入力データと標本画像の類似度計算だけで評価しているので、個々の動作の変わり目に他の口の状態を認識することが多く、評価が非常に難しい。例えば、『頭』という部分のジェスチャにおいては、a → e → n → e → a → ···· → e → a というようにバラバラの状態遷移を観測するので、認識しにくく誤認識が増える。実際、9 ジェスチャを 7 人分、つまり 63 個のデータを入力したとき、入力データ 1 枚 1 枚の認識率は約 80%であったのに対して、一連のジェスチャでは認識率約 5%となり、急激に認識率が低下した。

5.3.2 認識実験 2

状態番号は、『あ』が『0』、『い』が『1』、『う』が『2』、『え』が『3』、『お』が『4』、『ん』が『5』である。また、SAMPLE1～SAMPLE4 は HMM 学習に使用した画像で、SAMPLE5～SAMPLE7 は HMM 学習未使用画像である。

認識状況と認識率を表 2 に示す。表の○印は主語、述語の両方とも認識できたときである。△印は主語、述語の片方だけ認識できたとき、×印は両方とも認識できなかったときである。認識率算出においては、△も×と考えている。表 2 からわかるように実際に HMM 学習に使用したサンプルを入力データとした場合は認識率 86%という高い認識率を得られた。

しかし、HMM 学習未使用画像を入力データとした場合は認識率が 48%とかなり低下していることがわかる。

その理由の 1 つとして、SOM 学習において標本画像を作る際に、適当な画像を用いたことに原因があると考えられる。それにより、実際に学習データを入力したときに選ばれない標本が現れ、未学習データ入力時の認識率低下を招いていると考えられる。

	SAMPLE(学習使用)				認識率	SAMPLE(学習未使用)			認識率	TOTAL
	1	2	3	4	[%]	5	6	7	[%]	[%]
頭いたい	○	○	○	○	100	○	○	○	100	100
頭おもい	○	△	○	○	75	△	△	○	33	54
胸いたい	○	○	○	○	100	○	△	○	67	84
胸くるしい	○	○	○	△	75	○	×	○	67	71
お腹いたい	○	○	○	○	100	○	○	○	100	100
お腹すいた	○	○	△	○	75	△	△	△	0	38
背中いたい	○	○	△	○	75	△	△	○	33	54
手足いたい	○	○	○	○	100	△	△	○	33	67
手足しびれる	○	○	○	△	75	×	×	△	0	38
TOTAL [%]	100	89	78	78	86	44	22	78	48	67

表 2：認識状況

6 結論

本研究では、ジェスチャ認識の一つとしてリップリーディング・システムを試作した。認識の手法としては、ベクトル量子化を利用した自己組織マップ (SOM) と確率的オートマトンである隠れマルコフモデル (HMM) を用いた。

認識実験2において、パラメータ推定に用いたサンプルを入力データとした場合は、認識率 86%という高い認識率がえられ、HMM の有効性が確認された。しかし、HMM 学習未使用画像を入力データとした場合は認識率 48%とかなり低下した。これは、HMM 学習の際に現れなかった観測系列のデータが、学習未使用画像の入力に現れた場合の誤認識と考えられる。

今後の課題として、以下の点が挙げられる。

■今後の課題

- SOM よりも認識率が高いハイパーコラムモデル (HCM) などのシステムを用いた場合の認識率の検証
- 今回は観測系列の生成法として SOM を用いたが、実際にそれが有効だったかについてや、他に観測系列を生成する有効な手法はないかという点を検討する。
- 単語と単語の変わり目の誤認識を減らすために、今回は独自のパラメータ推定法を用いたが、HMM 学習において多量の学習サンプルを扱える Baum-Weich アルゴリズムを用いてパラメータ推定法や他の推定法について検討する。

参考文献

[1] 北 研二: “確率的言語モデル”, 東京大学出版会, 1999.

[2] 清水 宏明、岩井 儀雄、谷内田 正彦 : “HMM を利用したジェスチャ認識の高性能化”, 情報処理学会研究報告, pp. 105-112, 1999.

[3] 森山 慶: “自己組織マップを用いた動作予測に関する研究 –ジャンケンポン・システムの試作–”, 福岡大学工学部卒業論文, 2000.

[4] N. Tsuruta, R. Taniguchi, M. Amamiya, Proc. of 4th Work-Conf. on Aritificial and Natural Neural Networks (IWANN) (1999) 840 – 849

[5] N. Tsuruta, T. E. Tobely, Y. Yoshiki, Proc. of 6th Work-Conf. on Aritificial and Natural

Neural Networks (IWANN) (2001) to be published

[6] T. El. Tobely, Y. Yoshiki, R. Tsuda, N. Tsuruta and Makoto Amamiya, 6th International Conference on Soft Computing. IIZUKA2000 (2000) 207 - 214