コンピュータビジョンと
イメージメディア 103－15
（1997. 1. 24）

KL 展開と
隠れマルコフモデルによるジェスチャ認識

畠 直志　岩井 儀雄　谷内田 正彦

大阪大学 基礎工学部 システム工学科

Abstract

本論文では，接触型センサーやマーカーなどを装着することなく，また背景や人物が異なった場合などの環境の変化にもよらないロバストな人物の行動を認識するシステムを述べる．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に注目し，動領域を抽出することで，その動き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モーションベクトルを算出しKL 展開することでジェスチャ空間を作成する．そして，ジェスチャ空間に投影した各主成分を特徴量とし，それらを量子化することでシンボル化し，その時系列を HMM により学習する．ジェスチャの認識は，各 HMM 内での尤度を最大にするものを選択することで行なう.HMM による認識は，雑音や時間的な伸縮に対してロバストであり，学習に基づいて認識系が構築できる点で有効な手法である．

Gesture Recognition Combining HMM and KL Transform

Tadashi HATA,　Yoshio IWAI,　Masahiko YACHIDA

Department of Systems Engineering, Faculty of Engineering Science,
Osaka University

Abstract

This paper proposes a method to recognize human gestures from an image sequence based on Hidden Markov Model and Karhunen-Loève Transform. As our method uses the motion vector field of the scene for recognition, it is robust for variety in the background of the scene and it doesn't require the users to wear a sensor or a marker. The motion vector field of the scene is projected to an eigen-subspace for data compression and is used as the input symbols for the HMM.

1 はじめに

日常社会において, 人間は知らず知らずの内にジェスチャをし, 言語以上に意志を伝達することがよくある. そのジェスチャをコンピュータに認識, 理解させることは, 意志伝達をすみやかに行なうという点でも重要な課題である. 近年, マンマシンインターフェイス, ヒューマンマシンインタラクティブの研究分野において, コンピュータが人間の動作, ジェスチャ, 表情などを理解する研究が盛んに行なわれ, インターラクティブな意志の疎通の実現を目指した研究が進んでいる.

従来のジェスチャ認識の研究の多くは, 体にセンサやマーカーを付けたり, 背景, 個人差, 照明条件などの環境条件や画像の正規化を必要とするなど何らかの制約が必要となる. しかしながら, コンピュータとのよりよいインターフェイスを考えた場合, これらの制約を最小限におさえ, 環境条件によらないロバストな認識系の構築が望まれる.

本研究の関連例としては,DP マッチングを用いたジェスチャ動画像のスポッティング認識 [1] がある. これは, 動画像の時空間エッジ特徴を DP マッチングすることによりジェスチャ認識を行なった研究である.

また HMM を用いた研究にカテゴリー VQ を用いた動作認識法 [3] がある. これは, 特徴量にメッシュを用い, ベクトル量子化することでコードブックを作成し,HMM で認識を行なった研究である.

KL を用いた認識にインタラクティブシステム構築のための動画像からの実時間ジェスチャ認識法 [4] がある. これは, 改良したテンプレートモデルを用いて特定部位を抽出し, その画素値を KL 展開することでジェスチャ曲線として表現しモデル曲線と比較することによりジェスチャ認識を行なうものである.

本報告は, 6 つの章から構成されている. 次章では, 本手法の概要について, 第 3 章では, 動画像からの動き情報の抽出と KL 展開による情報圧縮と特徴抽出法を, 第 4 章では, 特徴量のシンボル化と HMM による認識手法を, 第 5 章では, 楽器を演奏するジェスチャをサンプルとした本手法による実験結果, 及び考察を, 最後の第 6 章では, 本論文の手法のまとめについて論じる.

2 本手法の概要

本研究では, 接触型センサーやマーカーなどを装着することなく, また背景や人物が異なった場合などの環境の変化にもよらないロバストな人物の行動を認識するシステムを提案する. 図 1 に, 本手法の簡略した過程を示す.

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得点に対して, 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.HMM による認識は, 雑音や時間的な伸縮に対してロバストであり, 学習にもとづいて認識系が構築できる点で有効な手法である.

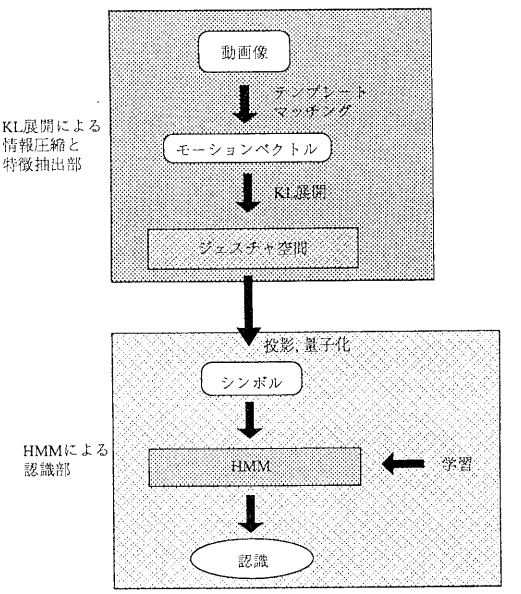

図 1: システムの過程

3 KL展開によるジェスチャ空間の構築

3.1 人物の動領域の発見

人物によるジェスチャ認識をするためには, まず最初に動画像中の情報から人物を発見し, 切り出す必要があるが, 容易なことではない.. そこでまず顔を発見することで, 画像中の人物の顔の位置を抽出 [5] し, それを基準に人物の動領域の切り出しを行なう. 図2は, 顔の抽出による動き領域の抽出図である.

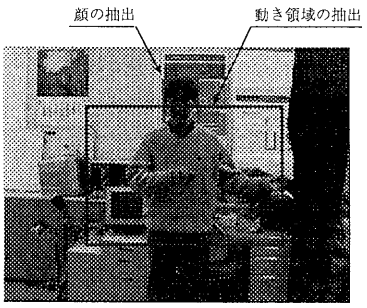

図 2: 顔の抽出による人物の動領域の抽出

3.2 動画像からのフロー抽出

動画像中から得られる画素をそのまま特徴量として用いると, 背景や照明条件といった環境の変化や人物の服装や個人差などの影響が大きくロバストな認識系を構築することができない.

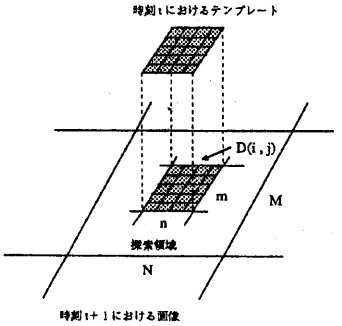

図 3: テンプレートマッチング

そこで, 人物の動き情報を特徴量とすれば, これらの環境の変化に強い認識系が構築できる. その動き情報としてオプティカルフローを用いる. オプティカルフローは, 物体の移動量と方向を表したもので, これを連続画像中でどの部分が互いに対応しているかを見つけることで求められる. このオプティカルフローはを抽出するための手法として, 図3のようなテンプレートマッチング法を用いる. テンプレートマッチング法は, テンプレート画像と探索画像の画素レベルでのマッチングを行なう相関法である.

時刻 t において位置 (m,n) のテンプレート画像の輝度を $x_t(m,n)$ とし, 時刻 t + 1 における位置 (m,n) の探索画像の輝度を $X_{t+1}(m,n)$ とすれば, その相違度である $D(i,j)$ は,

$$D(i,j) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} |X_{t+1}(i+m, j+n) - x_t(m,n)| \quad (1)$$

を計算することで求め, その最小相違度である D_{min} を

$$D_{min} = \min_{0\le i\le M-m, 0\le j\le N-n} D(i,j) \quad (2)$$

とすれば, これが $x_t(m,n)$ に最も類似するので, その x 軸方向と y 軸方向の移動量 (i,j) をモーションベクトルとする. このモーションベクトルは, 画像中の顔を基準とした $I \times J$サイズのウインドウを切り出し, 更に $M \times N$のブロックに分割し, その中でそれぞれ計算することにより求める. また, テンプレートマッチング法は, 計算コストが大きいことと, 雑音に敏感であるという問題点があるが, 前者は専用のハードウェアーを用いることで, 後者は HMM を用いることで解決することができる.

図4は, テンプレートマッチング法によって抽出されたオプティカルフローである. 各ブロック内で求められたオプティカルフローは, ラスター走査をすることにより, 次式 (3) のようにならべる. また, 人物の移動領域は顔を発見することで検出されているので, 抽出されたモーションベクトルを (x_i, y_j) とすれば, これをこの領域間のみで求めることにより探索領域を狭くし, 同時にこのフレームにおける特徴量 X として,

$$X = (x_1, y_1, x_2, y_2, \cdots, x_i, y_i, \cdots, x_N, y_N)^T \quad (3)$$

とする. ここで,N の次元数は $I \times J$である.

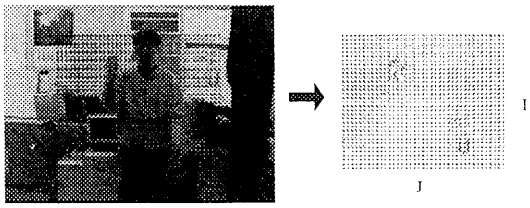

図 4: テンプレートマッチング法による抽出

3.3 KL 展開

動画像から得られたモデルとなる特徴量でパターン認識をするためには, その特徴量の相関を考え, できるだけモデルの特徴が反映するようにパターン空間を作成しなければならない. しかしながら, 得られた特徴量は次元数が非常に多く, 最適なパターン空間を作成するのは困難である. よって, 次に示すKL展開法を用いて低次元に落してパターン空間を作成する. KL展開は多変量の値をできるだけ情報の損失なしに, 少数個の総合的指標 (主成分) で代表させる多変量解析法 [4,6,7,8] である.

特徴量として得られたN次元ベクトル$X_n, (n = 1, 2, \cdots, N)$に対し, 正規直交ベクトル$\phi_k, (k = 1, 2, \cdots, K, (K \leq N))$とする. そして,$\phi_k$を決定するのに, 次の平均2乗誤差を考える. ここで,$y_k^{(n)} = (X_n, \phi_k)$とすると,

$$J(\phi_k) = \frac{1}{N}\sum_{n=1}^{N}\|X_n - \sum_{k=1}^{K} y_k^{(n)}\phi_k\|^2 \qquad (4)$$

この式が最小となるようにϕ_kを決定する. そこで,X_nの共分散行列をWとし,μをX_nの平均ベクトルとすると,

$$W = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu)(X_n - \mu)^T \qquad (5)$$

この共分散行列Wの固有ベクトルをϕとし, 対応する固有値をλとすると,

$$W\phi = \lambda\phi \qquad (6)$$

が成り立つ. この固有方程式を解いて得られる固有値を$\lambda_1, \lambda_2, \cdots, \lambda_k, (\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k)$とし, 対応する固有ベクトルを$\phi_1, \phi_2, \cdots, \phi_k$とすれば, これが正規直交ベクトルとなる.

また, 各主成分の分散が全体の中で占める割合である寄与率C_iは,

$$C_i = \frac{\lambda_i}{\sum_{j=1}^{K}\lambda_j} \qquad (7)$$

で与えられ, 第i固有ベクトルまでの累積寄与率SC_iは,

$$SC_i = \sum_{j=1}^{i} C_j \qquad (8)$$

で与えられる.

3.4 KL 展開による情報圧縮と特徴抽出

KL展開を画像に適用する場合に画素の輝度値を多変量の要素として画像の特徴抽出することが多いが, 本論文では, オプティカルフローを要素とする. 前に示したようにKL展開によって多次元の多数のモデルをパターン空間に記述することが可能となる. 具体的には,(3) 式の特徴量を各ジェスチャのモデル動作からそれぞれ得ることにより計N個を採用し,

$$X = (X_1, X_2, \cdots, X_N) \qquad (9)$$

とすると, これを先に述べた方法でKL展開して固有ベクトル

$$\phi = (\phi_1, \phi_2, \cdots, \phi_K) \qquad (10)$$

を得る. ここで, 固有値λ_iに対応する固有ベクトルϕ_iの要素は次式で表される.

$$\phi_i = (a_{i1}, a_{i2}, \cdots, a_{iM})^T \qquad (11)$$

但し,MはX_iの次元数に等価だから$M = 2 \times I \times J$である.

また, 第i固有ベクトルϕ_iの要素を係数とする合計変量をz_iとすれば,

$$z_i = \phi_i X = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{iM}x_M \qquad (12)$$

で表すことができ, これを第i主成分と呼び, その分散はλ_iである.

また, 主成分の数に関して, 複雑な現象をできるだけ単純化して理解しようという観点からいえば, 少ない数の主成分でもとの変量が代表されることが望ましい. 一般には, 累積寄与率が80%以上大きくなれば, ある程度代表されることになると知られている.

4 HMMによるジェスチャ認識

4.1 隠れマルコフモデル（HMM）

隠れマルコフモデル[10,11]は出力シンボルによって一意に状態遷移先が決まらないという意味での非決定性オートマトンとして定義される.出力シンボルが与えられても状態遷移系列は唯一に決まらず,観測できるのはシンボル系列だけであることから隠れマルコフモデルと呼ばれている.

隠れマルコフモデルは以下で定義される.

S: 状態の有限集合；$S = \{S_i\}$

Y: 出力シンボルの集合

π: 初期状態確率の集合；$\pi = \{\pi_i\}$

π_iは初期状態がS_iである確率

A: 状態遷移確率の集合；$A = \{a_{ij}\}$

a_{ij}は状態S_iから状態S_jへの遷移確率

B: 出力確率の集合；$B = \{b_{ij}(k)\}$

$b_{ij}(k)$は状態S_iから状態S_jへ遷移する際にシンボルkを出力する確率

4.2 主成分得点のシンボル化

離散HMMを用いて認識するには,ジェスチャ空間に投影した主成分をそのままHMMの入力シンボルとして扱うことができない.そこで,クラスタリング[9]を行なうことによって予めクラスターを作成しておき,その代表点である標準パターンとの比較によりシンボルに変換する.

すなわち,KL展開によって求められた固有ベクトルϕを(12)式により各主成分に対して計算すると,第i番目のモデルとなる特徴量Z_iは第K主成分までの

$$Z_i = (z_1, z_2, \cdots, z_K)^T \tag{13}$$

で表され,全てのモデル$(Z_1, Z_2, \cdots, Z_i, \cdots)$に対してクラスタリングを行なうことになる.

そして,主成分をシンボルに変換する際には,予めクラスターにラベル付けをしておき,入力主成分のベクトル$X = (x_1, x_2, \cdots, x_K)$と第iクラスターの代表点である標準ベクトル$Y_i = (y_{i1}, y_{i2}, \cdots, y_{iK})$とのユークリッド距離

$$d_i = \sqrt{\sum_{j=1}^{K}(x_j - y_{ij})^2} \tag{14}$$

を計算し,

$$d_{min} = \min_{1 \le i \le N} d_i \tag{15}$$

となるiをシンボルにすることによりシンボル化をする.これを各フレームに対して求めることによりシンボル系列を作成する.

4.3 HMMによるジェスチャ認識

HMMによる認識は,環境の違いなど雑音に対する影響や時間的な伸縮による影響を吸収することができる点で有効な認識手法である.また,HMMを用いた認識は,各ジェスチャiに対応するシンボル系列をHMMλ_iに予め学習させておき,観測されたシンボル系列$Y = y_1 y_2 \cdots y_T$がジェスチャiである確率$P(Y|\lambda_i)$が最大になるHMMを選択することによって行なう.

HMMによる学習は,$\{\pi$,A,B$\}$のHMMパラメータを推定することである.HMMのパラメータ推定には,Baum-Welchアルゴリズムという推定アルゴリズム[10,11]が確立されている.また,このパラメータを推定するには,多量の訓練用サンプルを要し計算量も多いが,このパラメータが収束するまで繰り返し学習を行なうことで推定を行なう.

5 実験及び考察

5.1 実験環境

入力画像のサイズは240x320とし,ウインドウの大きさは165x195とする.また,テンプレートは9x9,各探索エリアは27x27とし,5pixelごとにモーションベクトルを求めた.よって,各フレーム間のオプティカルフローの次元数は33x39=1287次元であり,x,y軸について求めるため総次元数は2574次元となる.

また,人物の動領域の発見は全画面をテンプレートマッチング法で探索していたのでは,コストが大きく,また人物が画面上の一定の位置にいないと行けないという制約ができてしまうので,ここでは,人物の顔を発見[5]し,その顔の大きさに合わせたウインドウを顔からの一定距離で切り出すことにより抽出する.これによって,人物が画面上のどこにいても,また,大きさが異なっていても安定して人物の動領域をとらえることができる.

図５は,KL展開によって固有空間上に表現したモデルジェスチャの寄与率と累積寄与率を表したものである.

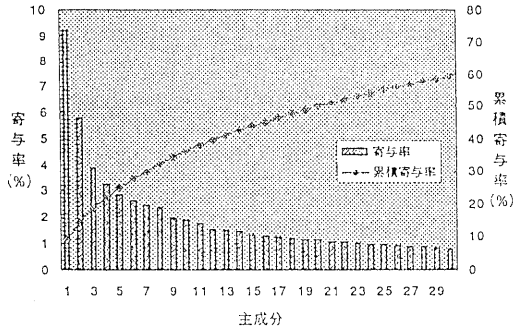

図 5: 寄与率と累積寄与率

認識に用いたジェスチャは, 楽器（ドラム, ギター, ピアノ, バイオリン, カスタネット）を演奏する５種類のジェスチャとする. ジェスチャ空間作成のためのモデル画像数は, 一つのジェスチャにつき６０枚, 合計３００枚用いた. そのサンプル画像を図６に示す. また,HMM の状態数は５とし, シンボル数を決定するクラスター数は３５とした.

5.2 個人差による実験

ジェスチャ空間に用いたモデル及びHMM の学習に用いたデータは同一人物によるもので, 認識にはその人物の他に４人のデータを使った. 先に述べた５種類のジェスチャに対してそれぞれ５０フレームの３つのサンプルを用意し, 合計１５個のデータを用いた. また, 主成分数は３０とした. 図７は, その実験結果である.

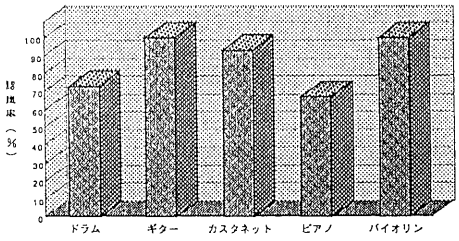

図 7: 各ジェスチャに対する認識率

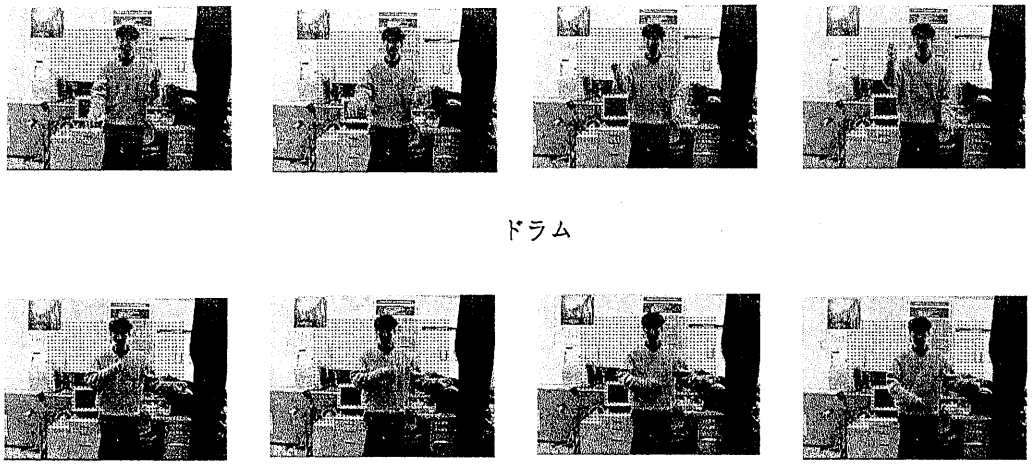

図 6: ジェスチャのサンプル画像

5.3 主成分数による実験

図5のグラフからも分かるように累積寄与率が50%を越えるのは第2 1主成分で,またある程度代表されるといわれる80%を越えるのは第6 9主成分となる.モデルを反映したデータを必要とするならここまでの主成分を使うべきであるが,出来るだけ少ないデータで認識が可能であるならば,それはHMMが吸収出来るという点で有効な手段である.よって,この実験では主成分数による認識実験を行なった.図8は,ドラムジェスチャの各主成分を軸とした第3主成分までの特徴量を時系列で表したものである.

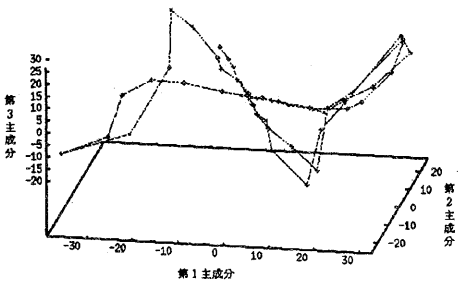

図 8: 第3主成分までの時系列データ

図9は他人物に対するドラムのジェスチャの主成分の数による認識率の推移例を表したものである.

学習に用いた人物が非学習の人物よりも認識率がよく,また認識できるフレーム数に対しても違いの差が大きく表れた.さらに,非学習の他人物が主成分数にかなり影響するのに比べて,学習者である人物ではほとんど影響がなかった.

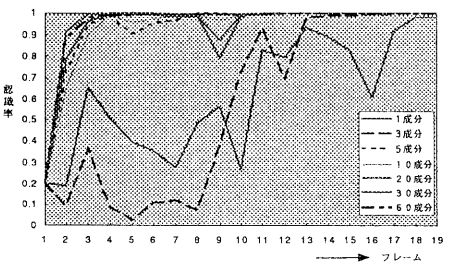

図 9: 他人物による主成分数に対する認識例

5.4 考察

図7を見ても分かるようにギター,カスタネット,バイオリンのジェスチャは100%の認識率であるが,ドラム,ピアノのジェスチャはあまり認識率が良くない.ピアノはドラムに,ドラムはカスタネットに間違うケースが多かった.モデルとして使った人物においては,いずれのジェスチャにおいても100%であったことからも分かるように,この2つのジェスチャにおいては個人差がかなり出てしまうようである.HMMを用いた認識の問題点としては,このような学習にないシンボル系列のデータを認識出来ない点があげられるので,この点を改善するには,まず,多数の人物によるジェスチャ空間の作成とHMMの学習を行なう必要がある.

また,主成分の数に関して,主成分を多く使うにつれ,認識率が向上し認識出来るフレームに差が出ると期待されたが,図9をみても分かるようにそのとおり大きく影響を及ぼした.しかしながら,学習者に対してはほとんど主成分の数による差がなく,第3主成分までだけで十分認識ができる結果になった.これによって,主成分を多数用いることで,他人物による異なるジェスチャに対して似たような系列が部分的に発生したとしても主成分を多数用いることで大幅に認識率を上げることができる.また,ジェスチャ空間の作成に関しても,モデルジェスチャを厳しく選択し有効と思われるようなものだけを固有空間に表現すれば,さらに認識率が向上すると思われる.

6 まとめ

本論文では,カメラ画像からの動き情報を抽出し,それを固有空間上に表現することによってジェスチャ空間を作成し,認識系にHMMを用いることによって,背景など環境の変化にロバストな人物のジェスチャ認識手法を述べた.今後は複数のカメラによる多眼視からの認識系を構築することによって,本手法の拡張性を検討していくつもりである.

参考文献

[1] 高橋勝彦, 関進, 岡隆一, "ジェスチャ動画像のスポッティング認識", 信学技法,PRU92-157, 1993.

[2] Junji YAMATO, Jun OHYA, Kenichiro ISHII, "Recognition Human Action in Time -Sequential Images using Hidden Markov Model", CVPR, pp.379-385, 1992.

[3] 大和淳司, 倉掛正治, 伴野明, 石井健一郎, "カテゴリー V Qを用いたHMMによる動作認識法", 信学論, Vol.J77-D-II, No.7, pp.1311-1318,1994.

[4] Takahiro WATANABE, Chil-Woo LEE, Masahiko YACHIDA, "Recognition of Complicated Gesture in Real-Time Interactive System", 5th IEEE International Workshop on RO-MAN, pp.268-273, 1996.

[5] Haiyuan. WU, Qian. CHEN, Masahiko. YACHIDA, "A Fuzzy-Theory-Based Face Detector", Proceeding of ICPR, Vol.3, No.13, pp.406-410, 1996.

[6] 赤松茂, 佐々木努, 深町映夫, 末永康仁, "濃淡画像マッチングによるロバストな正面顔の識別法", 電子情報通信学会,Vol.J76-D-II, No.7, pp.1363-1373, 1993.

[7] Hiroshi MURASE, Shree K. NAYAR, "Illumination planning for Object Recognition in Structure Environments", CVPR, pp.31-38, 1994.

[8] 舟久保登, "パターン認識", 共立出版,1991.

[9] Yoseph Linda, Andres Buzo, Rovert M.Gray, "An Alogorithm for Vector Quatizer Design", IEEE trans.Comnui, Vol.com-28, No.1, pp.84-94, 1980.

[10] 中川聖一, "確率モデルによる音声認識", 電子情報通信学会, 1988.

[11] Lawrence Rabiner, "A Tutorial on Hidden Marcov Models and Selected Applications in Speech Recognition", Proceeding of the IEEE, Vol.77, No.2, pp.257-285, 1989.