論　文

手書き数字認識における特徴選択に関する考察

若林　哲史[†]　　鶴岡　信治[††]　　木村　文隆[†]　　三宅　康二[†]

Study on Feature Selection in Handwritten Numeral Recognition

Tetsushi WAKABAYASHI[†], Shinji TSURUOKA[††], Fumitaka KIMURA[†], and Yasuji MIYAKE[†]

あらまし　本論文では，手書き数字認識において400次元の原特徴量を用い，判別分析法，主成分分析法を含むいくつかの特徴選択手法の有効性を計算機実験によって評価する．特徴選択手法の有効性は識別に用いる識別関数や識別手法に依存して異なるので，識別関数（手法）として，擬似ベイズ識別関数，2次識別関数，線形識別関数，投影距離，部分空間法を用いて特徴選択手法の評価・比較を行う．その結果，手書き数字認識における特徴選択の手法としては，主成分分析法が総合的に優れていることを示す．識別関数（手法）としては擬似ベイズ識別関数，投影距離・部分空間法が優れていることを示す．特に擬似ベイズ識別関数は特徴量の次元削減による識別率の低下が少なく，主成分分析法で特徴量を1/4程度にまで減らしても，正読率はほとんど低下せず，認識時間・記憶容量を節約することが可能となった．

キーワード　文字認識，数字認識，パターン認識，郵便番号認識，特徴抽出，特徴選択

1. まえがき

光学的文字認識（OCR）における手書き数字認識の研究の歴史は古いが，近年のOCRの普及と応用の広がりによって数字認識の高精度化・高速化の研究はより重要かつ活発になっている[1]～[5]．これは，郵便番号をはじめとして帳票類には数値項目が多いこと，単語認識，文章認識に比べて数字認識では，文脈依存の誤り検出・訂正が困難なことが理由として考えられる．また，文脈処理（知識処理）に基づく手書き文書解析の例として，郵便住所の解析・認識においても，膨大な住所データベースの効率的な検索に郵便番号や通り番号（street number）の数字認識が利用されており[6]，その重要性の高いことがわかる．

手書き数字認識を含め一般に文字認識の高精度化のためには，分離性の高い特徴量を抽出して用いる必要がある．分離性の高い特徴量を得る直接的な手段として，特徴抽出時における領域分割の細分化等により特徴量の次元数を増加させる方法が考えられる．

筆者らは文献[5]で特徴量の次元数の増加による手書き数字認識の高精度化に関する問題点について考察し，高次元特徴量（400次元）の利用により現時点で最高水準の正読率を達成した．しかし，次元数の増加により識別に要する計算量や記憶容量が増加する問題が発生する．この問題を解決するためには，正読率を低下させることなく特徴量の次元数を減少させる必要があり，統計的手法を用いて分離に適した特徴を選択することが有効である．また有効な次元削減を行えば，原特徴量を更に高次元化することによって正読率を向上させることも可能になる．

統計的手法による特徴選択手法としては，判別分析に基づく手法[7]が最も代表的であるが，字種数が少ない数字認識では識別力に関してソートできる特徴量が少なすぎる（9個）ため，特徴選択能力の低下を招く．この問題を解決するためにこれまで理論，実験の両面から多くの研究がなされている[8]．他の代表的な特徴選択手法として，主成分分析法と変数選択法がある[9]，[10]．文献[9]では，手書き文字の筆者識別において判別分析法と共に主成分分析法，変数選択法を用いて識別実験を行い，主成分分析法が特徴選択手法として有効であることを報告している．また文献[10]では，手書き漢字の類似文字認識において主成分分析法と変数選択法の改良手法が有効であることを報告している．

本論文では，手書き数字認識において原特徴量とし

† 三重大学工学部情報工学科，津市
†† 三重大学工学部電気電子工学科，津市
Faculty of Engineering, Mie University, Tsu-shi, 514 Japan

て400次元の特徴量を用い，判別分析法，主成分分析法を含むいくつかの選択手法の有効性を計算機実験によって評価する．また，特徴選択手法の有効性は識別に用いる識別関数や識別手法に依存して異なるので，識別関数（手法）として，擬似ベイズ識別関数，2次識別関数，線形識別関数，投影距離，部分空間法を用いて特徴選択手法の評価・比較を行う．

2. 特徴抽出

筆者らは文献[5]で，濃度値のこう配を特徴量とする手書き数字認識の有効性を報告した．本論文では，そのうち最も良い認識結果を示した400次元の特徴量を用いて特徴選択に関する実験を行う．以下に，画像の周辺分布と文字の位置情報を用いて切出しを行った後の数字データに対する特徴抽出手順を簡単に示す．

（1） 前処理として，外接枠・重心合せによる位置・大きさの正規化を行う．

（2） 全画素に2×2の平均値フィルタを施す処理をp回行うことにより濃度値画像を得る．文献[5]の結果から，$p=5$の場合に最も良い正読率が得られることがわかっている．

（3） 濃度値画像の濃度値の平均が0，分散が1となるように画像を正準化する．更に正準化画像に対してRobertsフィルタを適用し，こう配の向きと強さを求める．

（4） こう配の方向を$\pi/16$刻みで（向きの違いを考慮して）32方向に量子化する．

（5） 文字の外接枠を9×9の81個の小領域に分割し，各領域内において（4）で量子化した32の向き別にこう配の強さを加算することで局所方向ヒストグラムを作成する．

（6） [1 4 6 4 1]の加重フィルタにより，32方向のヒストグラムを，16方向に縮小する．更に方向別に，重なりのある2次元ガウスフィルタを施して，領域数を9×9から5×5に削減して400次元の局所方向ヒストグラム特徴を得る．

（7） 変数変換（$y=x^{0.4}$）により特徴量の分布形を正規分布に近づける[7]．

（8） 判別分析法，主成分分析法，変数選択法を用いて，400次元の特徴量からn'次元の特徴量を選択する（n' = 3, 6, 9, 12, 20, 30, 64, 80, 100, 144, 196, 256, 400）．特徴選択の手法については次章に述べる．

3. 特徴選択

特徴抽出理論の分野では，原特徴量から識別に有効な少数の特徴量を選択することを特徴選択，原特徴量のすべてを用いて識別に有効な少数の特徴を新たに生成することを特徴抽出と呼んでいる．しかし，本論文では前章の特徴抽出との混同を避けるため，この両者を特徴選択と呼ぶことにする．

比較に用いる特徴選択手法としては，分散比（F比）最大化の代表的手法である正準判別分析法と全クラスの混合分布を最小2乗誤差近似する主成分分析法（K-L展開）を用いる．また，比較のために，原特徴量の一部の変数を選択する変数選択法も用いる．

3.1 正準判別分析による方法（判別分析法）

n次元の確率変数ベクトルXに対し，級内分散行列S_w，級間分散行列S_bを次式で定義する．

$$S_w=\sum_{l=1}^{L}P(\omega_l)E\left\{(X-M_l)(X-M_l)^T\mid\omega_l\right\}=\sum_{l=1}^{L}P(\omega_l)\Sigma_l \quad (1)$$

$$S_b=\sum_{l=1}^{L}P(\omega_l)(M_l-M_0)(M_l-M_0)^T \quad (2)$$

但し，Lはクラス数，$P(\omega_l)$はクラスω_lの事前確率を示す．M_0はクラスごとの平均ベクトルM_lを用いて次式で定義する．

$$M_0=\sum_{l=1}^{L}P(\omega_l)M_l \quad (3)$$

ここで，式（4）を満たす固有ベクトル行列Φと固有値行列Λを求める．

$$S_b\Phi=S_w\Phi\Lambda \quad (4)$$

S_bのランクをrとすると$r=\min(L-1, n)$であり，最大r個の0でない固有値が得られる．Λの対角要素をλ_i（$\lambda_1\geqq\lambda_2\geqq\cdots\geqq\lambda_r$）とし，対応する固有ベクトルを$\Phi_i$（$\Phi=[\Phi_1\,\Phi_2\cdots\Phi_r]$）とする．このとき線形結合により得られる$z_i=\Phi_i^TX$，（$i=1, 2, \cdots, r$）を正準判別変量と呼ぶ．固有値の大きい方から$n'$個の正準判別変量を選択し，$n'$次元の特徴ベクトル$Z=(z_1, z_2, \cdots, z_{n'})$とする．

3.2 主成分分析による方法（主成分分析法）

3.2.1 全分散（固有値）の大きい順に主成分を選択

全共分散行列$S_t(=S_w+S_b)$を用いて次式を満たす固有ベクトル行列Φと固有値行列Λを求める．

$$S_t\Phi = \Phi\Lambda \qquad (5)$$

Λの対角要素をλ_i（$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n$），対応する固有ベクトルを$\Phi_i$（$\Phi = [\Phi_1\ \Phi_2 \cdots \Phi_n]$）とすると線形結合により$n$個の主成分$z_i = \Phi_i^T X$,（$i = 1, 2, \cdots, n$）が得られる．固有値の大きい方から$n'$個の主成分を選択し，$n'$次元のベクトル$Z = (z_1, z_2, \cdots, z_{n'})$とする．

3.2.2 F比の大きい順に主成分を選択

3.2.1で求められる，各主成分z_i（$i = 1, 2, \cdots, n$）に対し，級内分散s_{wi}^2と級間分散s_{bi}^2を求め次式によりF比を計算する．

$$F_i = \frac{s_{bi}^2}{s_{wi}^2} \qquad (6)$$

但し，

$$s_{wi}^2 = \sum_{l=1}^{L} P(\omega_l)\, E\left\{(z_i - m_{li})^2 | \omega_l\right\}$$

$$s_{bi}^2 = \sum_{l=1}^{L} P(\omega_l)\,(m_{li} - m_{0i})^2$$

$$m_{li} = E\{z_i | \omega_l\}$$

$$m_{0i} = E\{z_i\} = \sum_{l=1}^{L} P(\omega_l)\, m_{li}$$

z_iの中からF比の大きい順にn'個の主成分を選び，特徴ベクトル$Z = (z_1, z_2, \cdots, z_{n'})$を得る．

一般に，F比の大きな特徴の方が識別により有効であり，個々の特徴としての識別能力も高い傾向がある．従ってF比の大きな成分を選択した方が，全分散の大きな主成分を選択するより合理的と考えられる．なお，全分散λ_iは$s_{bi}^2 + s_{wi}^2$に等しい．

3.3 一部の変数を選択する方法（変数選択法）

元の確率変数ベクトル$X = (x_1, x_2, \cdots, x_n)$に対して，変数ごとにF比を計算し，F比の大きい順にn'個の変数を選ぶことでn'次元の特徴ベクトル$Z = (z_1, z_2, \cdots, z_{n'})$を得る．F比の計算は式（6）の$z_i$を$x_i$におき換えて行うものとする．

なお，S_w，S_b，S_t等の計算に用いる確率分布の各パラメータは学習標本から推定したものを用いる．また事前確率は，学習サンプルにおける総サンプル数をN，クラスのサンプル数をN_lとし，$P(\omega_l) = N_l / N$として計算した．

数字のみの認識ではクラス数$L = 10$である．そのため正準判別分析を用いると，級間分散が0でない正準判別変量は9個しか選択できず，10個目以降はクラスごとの平均がすべて同じになってしまうが，他の手法との比較のために12次元，20次元，…，400次元の特徴量に対する実験も行った．

4. 識別関数

特徴選択により作成した特徴量の識別力，あるいは分離性の良さを評価するため，代表的識別関数を用いて認識実験を行い，誤り確率を評価する．

識別関数としては，文献[5]で用いた擬似ベイズ識別関数以外に，特徴量の次元数と正読率の関係等が理論的によく研究されている2次識別関数と線形識別関数を用いる．また，擬似ベイズ識別関数と性質が近く，特徴抽出理論の分野でよく研究されている投影距離・部分空間法を用いる．

4.1 2次識別関数

パラメータが既知の正規分布に対するベイズ識別関数としてよく知られる2次識別関数は次式で表される．

$$g_l(X) = (X - M_l)^T \Sigma_l^{-1} (X - M_l) + \ln\left|\Sigma_l\right| - 2\ln P(\omega_l) \qquad (7)$$

しかし，現実のパターン認識では母集団の平均，共分散行列が既知であることはまれで，推定量として標本平均，標本共分散が用いられる．そのため特徴量の次元数を増加させると，共分散行列の推定誤差の影響で，ある次元を境に正読率が低下する尖頭現象（ピーキング現象）が生じる[11].

4.2 疑似ベイズ識別関数

共分散行列の推定誤差に起因する問題を解決するには，共分散行列が未知の正規分布に対するベイズ識別関数を用いるとよい．このベイズ識別関数は次式で与えられることが示されている[12].

$$g(X) = (N + N_0 + n - 1)\ln\left\{1 + \frac{(X - M)^T \Sigma_N^{-1} (X - M)}{N + N_0}\right\} + \ln\left|\Sigma_N\right| - 2\ln P(\omega) \qquad (8)$$

但し，

$$\Sigma_N = (1 - \alpha)\hat{\Sigma} + \alpha\Sigma_0$$

$$\alpha = \frac{N_0}{N + N_0}$$

ここで，nは特徴量の次元数，Nは各クラスの学習サンプル数，Mは平均ベクトル，$\hat{\Sigma}$は標本共分散行列，

Σ_0は母集団の共分散行列の初期推定量で，N_0はΣ_0の信頼度定数である．なおクラスに関する添字lは省略してある．

実験では，等方的初期分布を仮定して$\Sigma_0=\sigma^2 I$とし，計算量を削減するために次式を用いた．

$$g(X)=(N+N_0+n-1)\ln\left[1+\frac{1}{N_0\sigma^2}\left[\|X-M\|^2-\sum_{i=1}^{k}\frac{\lambda_i}{\lambda_i+\frac{N_0}{N}\sigma^2}\{\Phi_i^T(X-M)\}^2\right]\right]+\sum_{i=1}^{k}\ln\left(\lambda_i+\frac{N_0}{N}\sigma^2\right)-2\ln P(\omega) \quad (9)$$

式（8）から式（9）への変形は付録に付す．

4.3 投影距離・部分空間法

前述の疑似ベイズ識別関数において，$|\Sigma_N|$，$P(\omega)$がすべてのクラスで等しく，$i\leq k$の場合に$\lambda_i \gg (N_0/N)\sigma^2$を仮定すると，次式の投影距離[13]が得られる．

$$g(X)=\|X-M\|^2-\sum_{i=1}^{k}\{\Phi_i^T(X-M)\}^2=\sum_{i=k+1}^{n}\{\Phi_i^T(X-M)\}^2 \quad (10)$$

式（9）から，式（10）への変形は付録に付す．

この識別関数は特徴ベクトルXを共分散行列の最初のk個の主要固有ベクトルでK-L展開したときの2乗誤差であり，Xからこれらの固有ベクトルで張られる超平面への距離の2乗である．この超平面は，各クラスのサンプルの分布を近似する最小2乗誤差超平面となっている．

また，Xの属すクラスがXのノルム$\|X\|$に依存しない場合，$\|X\|=1$と正規化できる．更に$-X$の属すクラスがXの属すクラスに等しいと考えると$M=0$となる．この場合式（10）は次式のようになる．

$$g(X)=1-\sum_{i=1}^{k}\{\Phi_i^T X\}^2 \quad (11)$$

式（11）を最小化する決定則は，第2項を最大化する部分空間法[14]の決定則と等価である．

なお，疑似ベイズ識別関数，投影距離，部分空間法では，特徴ベクトルを各クラスで個別にK-L展開するのに対して，主成分分析では，特徴ベクトルを全クラスで共通にK-L展開する．このことから，部分空間法はクラスに固有の特徴選択の一手法と考えられている．また，擬似ベイズ識別関数，投影距離，部分空間法は，K-L展開によるクラスに固有の特徴選択能力を備えているため，共分散行列や自己相関行列の推定誤差の影響を受けにくく，必要な計算量や記憶容量も2次識別関数に比べて少ない．

4.4 線形識別関数

線形識別関数は，2次識別関数においてすべてのクラスの共分散行列が等しい（$\Sigma_l=\Sigma$）と仮定した場合に導かれる識別関数で，次式で表される．

$$g_l(X)=W_l^T X+W_{l0} \quad (12)$$

但し，

$$W_l=\Sigma^{-1}M_l$$

$$W_{l0}=-\frac{1}{2}M_l^T\Sigma^{-1}M_l+\log P(\omega_l)$$

式（12）の値は線形結合により求めることができ，決定境界は超平面となる．線形識別関数は識別に関係する未知パラメータの数が少ないため，パラメータの推定誤差の影響を受けにくい．また必要な計算量や記憶容量が少ないため，高速性を要求される識別や大分類に適している．

5. 認識実験

5.1 実験サンプル

実験には，郵政省郵政研究所が作成した第1回および第2回文字認識技術コンテストの学習用・評価用サンプルを用いる[5]，[15]，[16]．このサンプルは実際の郵便番号から収集された3けたの数字からなり，総サンプル数は14,954（44,862文字）である．

画像サイズは横240ドット×縦120ドットで，あらかじめ適当なしきい値で2値化されており，郵便番号枠は含まれない．筆記具は，ペン，ボールペン，サインペン，毛筆など多岐にわたっている．図1にサンプルの例を示す．

総サンプル数の内訳は次のとおりである．

（a） No.1～No.4979（4979サンプル）
　　第1回評価用，第2回学習用

（b） No.4980～No.9961（4982サンプル）
　　第1回評価用，第2回評価用

（c） No.9962～No.14954（4993サンプル）
　　第2回評価用

以後の認識実験では，（a），（b）を学習用サンプ

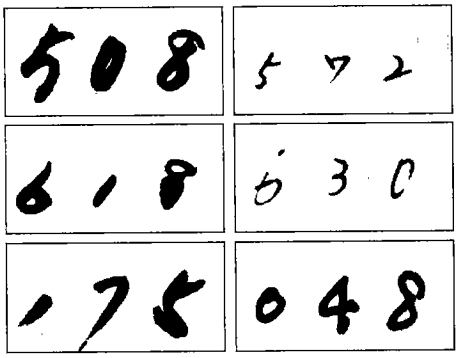

図1　テストサンプル
Fig. 1　Test samples.

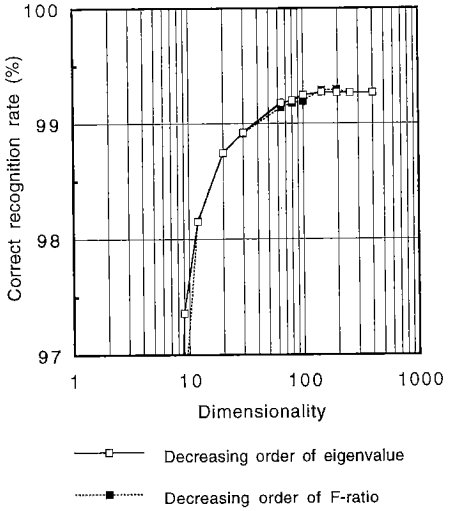

図2　次元数と正読率の関係（特徴選択：主成分分析法，識別関数：擬似ベイズ識別関数）
Fig. 2　Correct recognition rate v.s. dimensionality (feature selection: principal component analysis, discriminant function: pseudo Bayes).

ル，（c）を評価用サンプルとして用いる．文字数はそれぞれ29,883，および14,979である．

5.2 実験方法

3. で述べた4種類の特徴選択法で抽出した特徴量に対し，**4.** で述べた五つの識別関数を用いて認識実験を行い，評価用サンプルの文字ごとの平均正読率を求める．

なお，識別関数の各パラメータは学習標本から推定したものを用い，クラスの事前確率は等確率（$P(\omega)=1/L$）とした．

擬似ベイズ識別関数で用いる固有ベクトルは64次元以上の特徴量では37軸（$k=37$），30次元以下では次元数と同じ（$k=n$）とした．σ^2には全字種全固有値（分散）の平均を用いた．N_0は次元数により変化させるものとし，$\alpha=N_0/(N+N_0)$が0.1～0.9となる範囲で実験的に最も良い正読率を示すように選んだ．

投影距離・部分空間法で用いる固有ベクトルの数kは，第1軸から第k軸までの固有値の累積値が全固有値の和の80～90％程度となるようにクラスごとに選択する方法が一般的である．しかし予備実験の結果，クラス間でkを一定にする方が正読率が高いことがわかったため，実験では後者の手法を採用した．但しkは次元数により変化させるものとし，$1 \leqq k \leqq n-1$の範囲で，最も良い正読率を示すように選んだ．

更に比較のために，特徴選択を行わず，濃度値こう配を用いて，直接低次元の特徴量を求め，認識実験を行った（特徴選択無し）．この場合には局所方向ヒストグラムを作成する際に，方向量子化数・小領域分割数を変化させることで，9種類の次元数の特徴量を求めることができる（n = 16，36，64，80，100，144，196，256，400）．文献[5]の結果により，144次元以上では向きの違いを考慮した縮小16方向の特徴量を，100次元以下では向きの違いを考慮しない縮小4方向の特徴量を用いた．縮小16方向の特徴量は，32方向に量子化して求めた局所方向ヒストグラムを加重フィルタにより16方向に縮小したもの，縮小4方向は同様に8方向の特徴量を4方向に縮小したものである．

5.3 実験結果と考察

主成分分析法で選択した特徴量を用いて，擬似ベイズ識別関数による認識実験を行った結果を図2に示す．横軸は特徴選択後の次元数（対数表示），縦軸は正読率を表す．i) 全分散の大きい順に主成分を選択した場合と，ii) F比の大きい順に主成分を選択した場合を比較すると，64次元～100次元では前者が，144次元，196次元では後者がいくぶん上回るものの，両者の正読率のグラフには大きな差は認められない．従って以後主成分分析法の結果は，より一般的かつ処理が単純な i) のみ示すことにする．手法 ii) がより合理的と考えられるにもかかわらず，両手法に有為な差がなかった理由として，実験に用いた原特徴量の主成分の選択には，級内分散より級間分散が大きな役割を果たしていることが予想される．例えばすべての主成分の級内分散が等しい特殊な場合，手法 i) と手法 ii) は同一結果を与える．

各々の特徴選択法で擬似ベイズ識別関数による認識実験を行った結果を図3に示す．単純選択法の正読率は，144次元以上では特徴選択なしと同等であるが，100次元以下では下回っている．判別分析法の正読率に注目すると，9次元までは他の方法に比べて高い値を示している．しかしそれ以上の次元数では，196次元，256次元で正読率が上がるものの，ほとんど横ばいの状態である．400次元すべてを選択しても特徴選択なしに及ばない．一方主成分分析法は，12次元までは判別分析法に及ばないが，30次元以上では最も良い正読率を示している．64次元でも99％以上の値を保っており，144次元では400次元よりやや良い結果になっている．特徴選択なしと比べ常に高い正読率を示しており，同じ次元数なら400次元の特徴量を求めてから，主成分分析法によって特徴選択する方が，認識精度が高いことがわかる．また，次元数を大きく減らして識別計算を高速化する場合には判別分析法が有利であるが，正読率を落とさずに次元数を1/4程度にまで減らすには，主成分分析法が有効であると言える．

図4に2次識別関数を用いた認識実験の結果を示す．共分散行列の推定誤差のために尖頭現象が生じ，判別分析法，主成分分析法，特徴選択なしの場合にそれぞれ，9次元，30次元，64次元において正読率が最高になっている．このことから，2次識別関数では正読率を向上させるために積極的に特徴選択を行い，最適な次元数に次元減少する必要があることがわかる．しかし，最適次元数における2次識別関数の正読率は，その次元における擬似ベイズ識別関数の正読率より低く，必要な計算量，記憶容量も多い．

2次識別関数の性能劣化の原因として学習サンプル数の不足が考えられるが，本実験では1字種平均約3,000文字（計29,883文字）を使用しているため，学習サンプル数の増加によって2次識別関数の性能劣化を防ぐのは容易ではないと考えられる．

2次識別関数の学習サンプル数は特徴量の次元数の2次のオーダで増加させる必要があることが指摘されている[11]，[17]．この指摘に従えば，主成分分析法を用いた場合の最適次元数を30から300に移すには，1字種平均300,000文字の学習サンプルが必要になる．

図5に線形識別関数を用いた認識実験の結果を示す．特徴選択なし，主成分分析法，判別分析法の順で正読率が高くなることが明らかである．判別分析法の正読率は9次元以降ほぼ一定で，判別分析法で擬似ベイズ識別関数を用いた場合に近い値を示している．

線形識別関数で，特徴選択なしの400次元の特徴量を用いて識別を行うには，式(12)より400次元の内積を10回計算すればよいことがわかる．しかし，特徴選択を行って9次元の特徴量を求めるには，400次元の

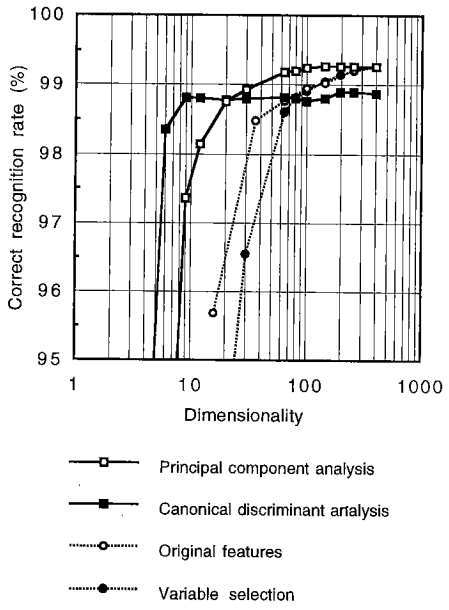

図3　次元数と正読率の関係（擬似ベイズ識別関数）
Fig. 3　Correct recognition rate v.s. dimensionality (pseudo Bayes discriminant function).

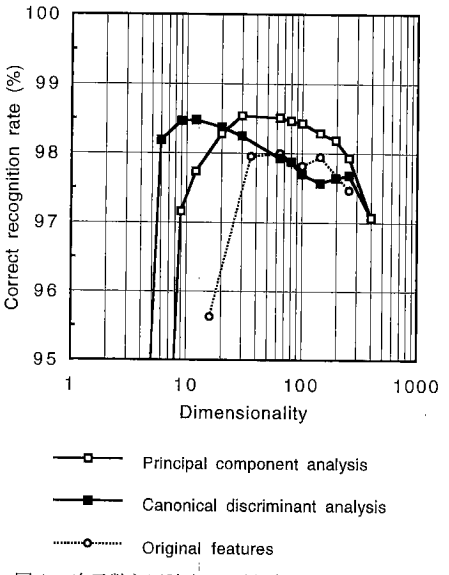

図4　次元数と正読率の関係（2次識別関数）
Fig. 4　Correct recognition rate v.s. dimensionality (quadratic discriminant function).

内積が9回必要となり，速度面での向上は期待できない．12次元以上では正読率が変わらないのにかえって計算量が増加してしまう．従って，線形識別関数では，特徴選択を行わずに400次元の特徴量を用いて認識を行うのが適していると考えられる．

図6，図7にそれぞれ投影距離，部分空間法を用いた認識実験の結果を示す．投影距離，部分空間法の正読率は，基本的に疑似ベイズ識別関数の正読率と同様の曲線になるが，特徴選択なしと主成分分析法を用いた場合には疑似ベイズ識別関数より多少低く，次元数が小さいほどその差は拡大している（図8，図10）．また判別分析法を用いた場合の投影距離，部分空間法の正読率は，高い次元では疑似ベイズ識別関数に近いが，次元数が小さくなると大きく低下し，2次識別関数や線形識別関数に及ばない（図9）．

低い次元数で投影距離，部分空間法の正読率が低下するのは，クラスごとの部分空間に共通部分（各クラスを近似する超平面が互いに交差する部分）が増加するのが原因と考えられる．特に判別分析法では，10軸以降で級間分散の固有値が0（つまりクラスごとの平均がすべて同じ）となるため，部分空間に必ず共通部分が発生し，正読率の低下を引き起こしていると考えられる．

共通部分空間とその近傍にあるサンプルに対する識別誤りが増加する部分空間法の欠点は，識別に固有値（分散）を用いていないために生じる問題である．部分空間法を改良してこの問題を解決する方法として，共通部分空間を除去する方法[18]等が検討されているが，十分な解決策は見出されていなかった．固有値を用いる疑似ベイズ識別関数では，この問題は生じない．図11は投影距離を用いた場合の決定境界の例で，

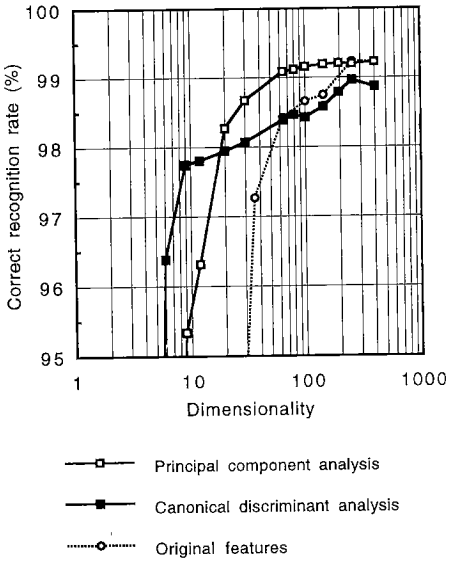

図6　次元数と正読率の関係（投影距離）
Fig. 6　Correct recognition rate v.s. dimensionality (Projection distance).

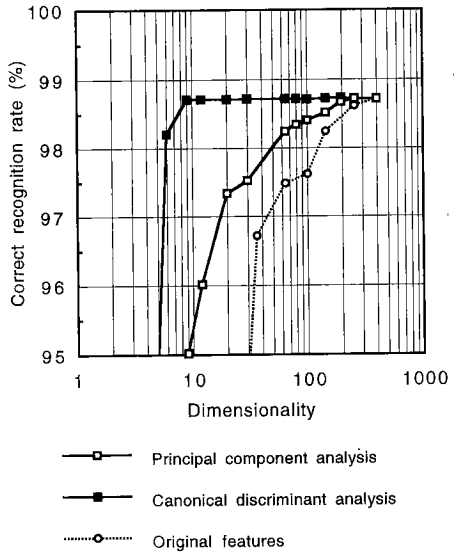

図5　次元数と正読率の関係（線形識別関数）
Fig. 5　Correct recognition rate v.s. dimensionality (linear discriminant function).

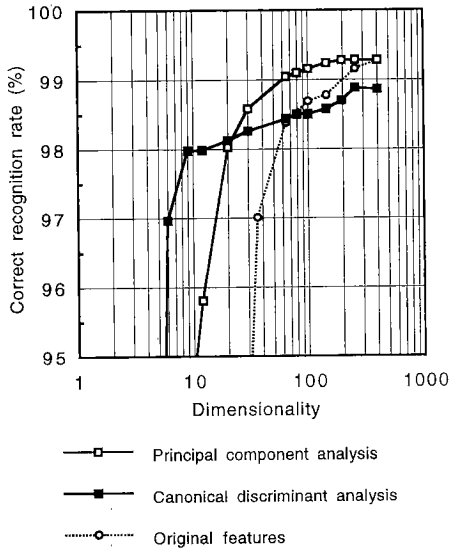

図7　次元数と正読率の関係（部分空間法）
Fig. 7　Correct recognition rate v.s. dimensionality (Subspace method).

灰色部分 $\Re(\omega_1)$ にあるサンプルはクラス1に分類される．この例では二つの分布の長軸が各クラスの部分空間で，超平面は直線となり，決定境界はこの直線からの等距離線となる．その結果，共通部分空間（2直線の交点）とその近傍にあるサンプルに対して，多くの識別誤りが生じている．一方図中の破線は擬似ベイズ識別関数を用いた場合の決定境界である．この場合，決定境界は固有値（分散）から定まる双曲線となり，上述の識別誤りは生じない．

図3，図6，図7に示す結果から，擬似ベイズ識別関数や投影距離・部分空間法を用いる手書き数字認識において，正読率を落とさずに特徴選択を行うには主成分分析法が有効であることがわかる．また図4，図5に示す結果では，2次識別関数や線形識別関数を用いる場合には判別分析法が有効であるが，2次識別関数の場合，学習サンプル数が増加すると主成分分析の方が有効となる可能性もある．

表1に，n 次元の原特徴から n' 次元の特徴選択を行った場合の認識に要する処理時間と記憶容量のオーダを，特徴選択と識別に分けて示す．特徴選択・識別ともに処理時間と記憶容量は特徴選択後の次元数 n' に比例しており，n' を小さくすることで認識システムの高

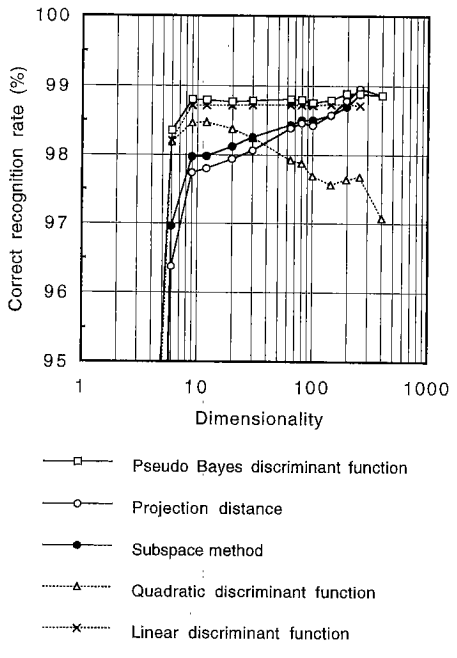

図9　次元数と正読率の関係（判別分析法）
Fig. 9　Correct recognition rate v.s. dimensionality (Canonical discriminant function).

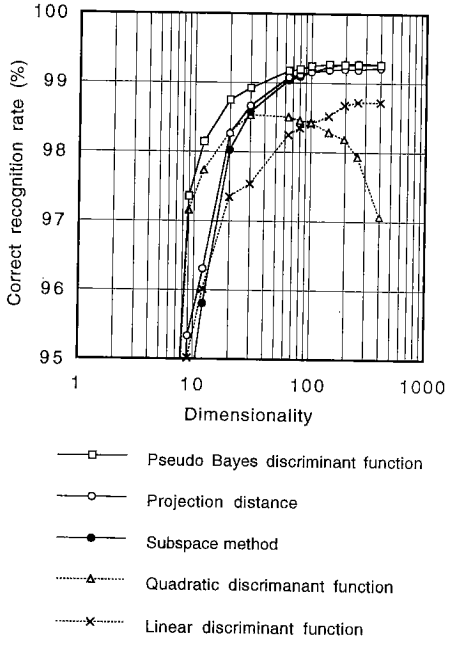

図8　次元数と正読率の関係（主成分分析法）
Fig. 8　Correct recognition rate v.s. dimensionality (Principal component analysis).

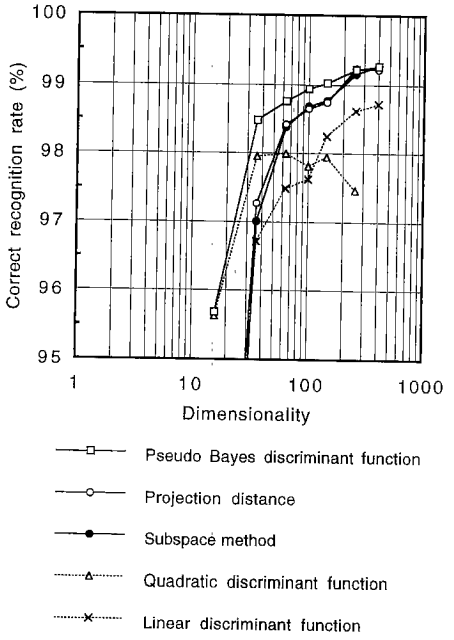

図10　次元数と正読率の関係（特徴選択なし）
Fig. 10　Correct recognition rate v.s. dimensionality (Original features).

速化・小容量化を図ることができる．識別においては，漢字のようにクラス数Lの大きい場合，識別に要する処理時間と記憶容量も一挙に増加するが，次元数n'の縮小による効果も大きくなる．kは疑似ベイズ識別関数，投影距離，部分空間法で用いる固有ベクトル数である．線形識別関数では$k=1$，2次識別関数では$k=n'$となる．

表2に特徴選択なしで400次元の特徴量を用いて認識を行う場合と，主成分分析法で144次元，100次元にしてから認識を行う場合の1文字当りの認識時間を示す．同表には，それぞれの識別辞書等の容量も示す．識別関数は，疑似ベイズ識別関数である．100次元の特徴選択をする場合400次元の内積計算が100回必要であるため，特徴抽出の時間が少し増えているが，次元数が減少したことで識別に要する時間が少なくなり，結果的に認識時間全体の縮小につながっていることがわかる．また，400次元の特徴量から100次元の特徴量を選択したことで，必要な記憶容量の合計が1/2に減少していることがわかる．正読率は144次元，100次元の場合にそれぞれ99.271％，99.245％となり，実際の郵便物から収集した手書き数字に対して現時点で最も良い結果が得られた．

表1　処理時間・記憶容量のオーダ
Table 1　Order of processing time and storage.

	処理時間		記憶容量	
	特徴選択	識別	特徴選択用変換行列	識別辞書
オーダ	$O(nn')$	$O(n'kL)$	$O(nn')$	$O(n'kL)$

6. むすび

手書き数字認識における特徴選択について，大量の文字サンプルを用いた実験で，選択した次元数と正読率の関係を調べた．

一連の実験において特徴選択の手法としては，主成分分析法が判別分析法より総合的に優れていた．その主な理由は判別分析法では，F比が0でない特徴量が9個（クラス数－1）しか得られないのに対して，主成分分析では一般にすべての特徴量が正のF比をもつからと考えられる．

識別関数（手法）としては擬似ベイズ識別関数，投影距離・部分空間法が2次識別関数，線形識別関数に比べて優れていた．特に擬似ベイズ識別関数は，クラス間の共通部分空間とその近傍で投影距離・部分空間法より識別力が高く，特徴量の次元削減による識別率

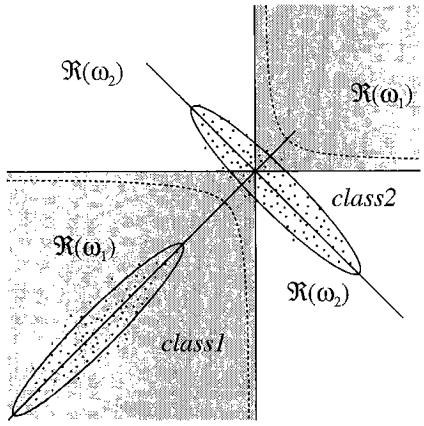

図11　投影距離による決定境界の例
Fig. 11　Example of decision boundary for projection distance.

表2　認識に要する処理時間・記憶容量の比較
Table 2　Comparison of processing time and storage.

	処理時間（ms／文字）				記憶容量（KB）			正読率(%)
	特徴抽出	特徴選択	識別	計	特徴選択用変換行列	識別辞書	計	
特徴選択無し 400次元	38.1	----	57.3	95.4	-----	609	609	99.265
特徴選択 144次元	38.1	16.3	20.7	75.1	226	219	445	99.271
特徴選択 100次元	38.1	11.1	14.4	63.6	157	152	309	99.245

の低下が少ない．

以上の理由から，手書き数字認識においては，主成分分析と擬似ベイズ識別関数を用いた場合に最も良い結果が得られ，特徴量を1/4程度にまで減らしても，正読率がほとんど低下せず，認識時間・記憶容量を節約することが可能となった．

少クラスの分類問題における判別分析法の制約をとり除くために，正規直交条件の下で分散比を最大化する方法[19]，[20]，共分散行列の差異を表す項を追加して特徴評価関数を拡張する方法[21]，[22]，共分散行列のノンパラメトリック化[23]などの方法が提案されている．今後の課題として，これらの特徴選択手法の比較や，識別手法としてニューラルネットワークや最近傍法を用いた場合の比較，漢字のようにクラス数が多い場合の検討，ニューラルネットワークによる特徴選択との比較等，興味のある問題が数多く残されている．

謝辞 本研究に関し，手書き郵便番号データを作成・提供された郵政省郵政研究所の皆様，討論をして頂いた研究室の皆様に感謝致します．

文　献

[1] T. Wakahara, "Handwritten Numeral Recognition using LAT with Structural Information," Proceedings of the Third International Workshop on Frontiers in Handwriting Recognition, pp.164-174, Buffalo, May 1993.

[2] G. Srikantan, "Gradient Representation for Handwritten Character Recognition", Proceedings of the Third International Workshop on Frontiers in Handwriting Recognition, pp.164-174, Buffalo, May 1993.

[3] H. Yan, "Digit Recognition", Proceedings of the Second International Conference on Document analysis and Recognition, pp.10-13, Tsukuba, 1993.

[4] T. Kawatani, "Handprinted Numeral Recognition with the Learning Quadratic Discriminant Function," Proceedings of the Second International Conference on Document analysis and Recognition, pp.14-17, Tsukuba, 1993.

[5] 若林哲史, 鶴岡信治, 木村文隆, 三宅康二, "特徴量の次元数増加による手書き数字認識の高精度化," 信学論（D-II）, vol. J77-D-II, no. 10, pp.2046-2053, Oct. 1994.

[6] S.N. Srihari, V. Govindaraju, and A. Shekhawat, "Interpretation of Handwritten Addresses in US Mailstream," Proceedings of the Second International Conference on Document analysis and Recognition, pp.291-294, Tsukuba, 1993.

[7] K. Fukunaga, "Introduction to Statistical Pattern Recognition, Second Edition," Academic Press, 1990.

[8] 浜本義彦, "パターン認識理論の最近の動向," 信学誌, vol.77, no.8, pp.853-864, Aug. 1994.

[9] 吉村ミツ, 木村文隆, 吉村　功, "わく内自由手書き片仮名の筆者識別法の比較," 信学論（D）, vol.J63-D, no.10, pp.819-826, Oct. 1980.

[10] 横塚志行, 阿部一朗, "類似カテゴリセットにおける特徴選択・次元圧縮手法の改良," 信学技報, PRU90-30, July 1990.

[11] K. Fukunaga and R.R. Hayes, "Effects of Sample Size in Classifier Design," IEEE trans. vol.PAMI-11, no.8, pp.873-885, Aug. 1989.

[12] F. Kimura and M. Shridhar, "Handwritten Numeral Recogniton based on Multiple Algorithms," Pattern Recognition, vol.24, no.10, pp.969-983, 1991.

[13] 池田正幸, 田中英彦, 岡本　達, "手書き文字認識における投影距離法," 情処学論, vol.24, no.1, pp.106-112, Jan. 1983.

[14] エルッキ・オヤ著, 小川英光, 佐藤　誠訳, "パターン認識と部分空間法," 産業図書, 1986.

[15] 松井俊弘, 山下郁生, 若原　徹, 吉室　誠, "文字認識アルゴリズム複合化の検討, 第一回文字認識技術コンテストの結果より," 信学技報, PRU92-33(1992-09).

[16] 能見　正, 松井俊弘, 山下郁生, 若原　徹, 吉室　誠, "手書き数字認識における誤読・リジェクトパターンの分析," 信学技報, PRU93-46, Sept. 1993.

[17] S. Raudys, "On Dimensionality, Learning Sample Size and Complexity of Classification Algorithms," Proc. Third Int. Joint Conf. Pattern Recognition, San Diego, pp.166-169, 1976.

[18] S. Watanabe and N. Pakvasa, "Subspace Method of Pattern Recognition," Proc. 1st Int. J. Conf. on Pattern Recognition, Washington DC, Oct.30-Nov.1, 1973.

[19] T. Okada and S. Tomita, "An Optimal Orthonormal System for Discriminant Analysis," Pattern Recognition, vol.18, no.2, pp.139-144, 1985.

[20] Y. Hamamoto, T. Kanaoka and S. Tomita, "On a Theoretical Comparison between the Orthonormal Discriminant Vector Method and Discriminant Analysis," Pattern Recognition, vol.26, no.12, pp.1863-1867, 1993.

[21] W. Malina, "On an Extended Fisher Criterion for Feature Selection," IEEE Trans, vol.PAMI-3, no.5, pp.611-614, Sept. 1981.

[22] M. Aladjem, "Parametric and Nonparametric Linear Mappings of Multidimensional Data," Proc. 11th Int. Conf. Pattern Recognition, The Hague, pp.101-104, 1992.

[23] K. Fukunaga and J.M. Mantock, "Nonparametric Discriminant Analysis," IEEE Trans, vol.PAMI-5, no.6, pp.671-678, Nov. 1983.

[24] 若林哲史, 鶴岡信治, 木村文隆, 三宅康二, "手書き文字認識における特徴量の次元数と変数変換に関する考察," 信学論（D-II）, vol.76-D-II, no.12, pp.2495-2503, Dec. 1993.

付　録

1. 擬似ベイズ識別関数の導出

文献[24]では，$N+N_0 \gg n$，$N \gg N_0$の仮定の下に疑似ベイズ識別関数を導出する手順を示したが，ここではNが十分大きくない場合の導出を簡単に示す．

式（8）において，

$$Y=(X-M)^T\Sigma_N^{-1}(X-M) \qquad (A\cdot 1)$$

とおく．$\Sigma_0=\sigma^2 I$を仮定して，

$$\Sigma_N=(1-\alpha)\hat{\Sigma}+\alpha\sigma^2 I \qquad (A\cdot 2)$$

とすると，式(A・1)は以下のように変形できる．

$$Y=\sum_{i=1}^{n}\frac{1}{(1-\alpha)\lambda_i+\alpha\sigma^2}\{\Phi_i^T(X-M)\}^2 \qquad (A\cdot 3)$$

$i>k$ の場合に $(1-\alpha)\lambda_i \ll \alpha\sigma^2$ を仮定すると次式の近似が得られる．

$$Y\approx\sum_{i=1}^{k}\frac{1}{(1-\alpha)\lambda_i+\alpha\sigma^2}\{\Phi_i^T(X-M)\}^2 + \sum_{i=k+1}^{n}\frac{1}{\alpha\sigma^2}\{\Phi_i^T(X-M)\}^2 \qquad (A\cdot 4)$$

以下の関係，

$$\sum_{i=k+1}^{n}\{\Phi_i^T(X-M)\}^2=\|X-M\|^2-\sum_{i=1}^{k}\{\Phi_i^T(X-M)\}^2 \qquad (A\cdot 5)$$

を用いると次式を得る．

$$Y\approx\frac{1}{\alpha\sigma^2}\left[\|X-M\|^2-\sum_{i=1}^{k}\frac{(1-\alpha)\lambda_i}{(1-\alpha)\lambda_i+\alpha\sigma^2}\{\Phi_i^T(X-M)\}^2\right] \qquad (A\cdot 6)$$

一方 $(1-\alpha)$ は正の定数であるから，$i>k$ で $\lambda_i \ll \alpha\sigma^2/(1-\alpha)$ とすると，

$$\begin{aligned}\ln|\Sigma_N| &= \sum_{i=1}^{n}\ln\{(1-\alpha)\lambda_i+\alpha\sigma^2\}\\ &= \sum_{i=1}^{n}\ln\left(\lambda_i+\frac{\alpha}{1-\alpha}\sigma^2\right)+\sum_{i=1}^{n}\ln(1-\alpha)\\ &\approx \sum_{i=1}^{k}\ln\left(\lambda_i+\frac{\alpha}{1-\alpha}\sigma^2\right)+\sum_{i=k+1}^{n}\ln\left(\frac{\alpha}{1-\alpha}\sigma^2\right)\\ &\quad+\sum_{i=1}^{n}\ln(1-\alpha)\end{aligned} \qquad (A\cdot 7)$$

であり，第2項，第3項は定数項である．

式（8）に式(A・6)(A・7)と $\alpha=N_0/(N+N_0)$ を代入し，定数項を省略すると式（9）の擬似ベイズ識別関数を得る．

2. 投影距離の導出

すべてのクラスで $|\Sigma_N|$，$P(\omega)$ が等しいと仮定すると，式(A・7)より擬似ベイズ識別関数（式（9））の第2項，第3項は定数項となる．また N はすべてのクラスで等しいため第1項の係数 $(N+N_0+n-1)$ はクラス間で共通である．更に，$i\leqq k$ において $\lambda_i \gg (N_0/N)\sigma^2$ を仮定し，定数項と第1項の係数を省略すると次式が得られる．

$$g(X)=\ln\left[1+\frac{1}{N_0\sigma^2}\left[\|X-M\|^2-\sum_{i=1}^{k}\{\Phi_i^T(X-M)\}^2\right]\right] \qquad (A\cdot 8)$$

ここで，

$$Z=\|X-M\|^2-\sum_{i=1}^{k}\{\Phi_i^T(X-M)\}^2$$

$$\beta=\frac{1}{N_0\sigma^2}$$

とおくと，$\ln(1+\beta Z)$ は Z に関する単調増加関数であるから，式(A・8)より式（10）の投影距離が得られる．

（平成6年11月10日受付，7年5月15日再受付）

若林　哲史（正員）

昭60三重大・工・電子卒．昭62同大大学院修士課程了．平2三重大・工・助手．現在に至る．手書き文字認識，文書理解，画像処理，コンピュータグラフィックスの研究に従事．

鶴岡　信治（正員）

昭52岐阜大・工・電子卒．昭54名大大学院博士（前期）課程了．同年三重大・工・電子助手，平1同大・工・助教授，現在に至る．この間，平3～4米国ミシガン大デアボーン校客員助教授．工博．手書き文字認識，文書理解，コンピュータグラフィックス，医用画像処理に関する研究に従事．情報処理学会，日本ＭＥ学会，人工知能学会各会員．

木村　文隆（正員）

昭48名大・工・電気卒．昭53同大大学院博士課程了．同年同大・工・助手，昭58三重大・工・助教授，現在に至る．その間平1からミシガン大学客員助教授を勤める．文字・パターン認識，画像処理，コンピュータグラフィックスの研究に従事．情報処理学会，日本ＭＥ学会，人工知能学会各会員．工博．

三宅 康二 （正員）

昭35名大・工・電気卒．昭40同大大学院博士課程了．同年同大・工・助手，昭43同講師，昭44同助教授，昭53三重大・工・教授，現在に至る．現在，手書き文字の機械認識，コンピュータグラフィックスおよび医用画像処理の研究に従事．情報処理学会，日本ＭＥ学会，日本人工知能学会，日本ロボット学会，電気学会，日本シミュレーション学会等各会員．工博．