手書き文字認識における複数特徴を統合する認識器EID3の提案

井藤好克[†,☆] 大橋 健[†] 江島俊朗[†]

今日まで手書き文字認識の分野では，多くの認識特徴が提案されてきたが，識別能力において決定的なものは存在しない．そこで本論文では，複数の認識特徴を統合することによる，認識性能の向上に関して論じる．既存の認識特徴を統合する手法は大きく2つに分類できる．1つは，順次に異なる認識特徴を適用して絞り込みを行う階層的分類手法．もう1つは，単一の認識特徴を用いた各識別器の出力を並列に統合して総合的な結果を得る並列分類手法である．階層的分類手法は，その設計において認識器設計者の調査が予測・試行錯誤などを必要とするうえ，認識特徴の追加が容易ではない．一方，並列分類手法は認識特徴の追加は容易であるが，識別に要する計算量は認識特徴の増加に比例して増大する．そこで，効率の良い統合を自動的に行う仕組みとしてEID3を提案した．EID3は木構造に識別器を配置し，未知パターンの入力に対しEID3上をパス選択して，識別に必要な統合すべき識別器を決定する．EID3は，特徴空間における各学習サンプルの分布状況を評価して，局所的にみて最適な識別器を設計し，識別のための決定木を漸次構築する．実験では，ETL6の手書き数字を対象に，EID3を含めた統合手法の統合効果について検証を行った．その結果，EID3では，既存の統合手法よりも少ない識別時間による統合効果を確認した．

EID3: An Integration of Multiple Features for Handwritten Characters Recognition

YOSHIKATSU ITO,[†,☆] TAKESHI OHASHI[†] and TOSHIAKI EJIMA[†]

The use of multiple features makes the performance of classification better. However, the design of a classifier with multiple features is difficult and often done by the designer's intuition and/or trial and error. In this paper, we propose a decision tree called EID3, which is automatically designed so as to minimize classification errors based on an evaluator of entropy. The EID3 is considered as an extension of well-known ID3 proposed by Quinlan etc and its construction is similar to that of ID3. Our experiment shows that the EID3 is a promising classifier for handwritten character recognition.

1. まえがき

手書き文字はパターン認識の研究対象の1つである．文書の中には手書きの文書も多く存在し，それらの文書情報を電子的に処理するためには手書き文字の符号化が必要になる．ここで高い認識能力を持つ手書き文字認識が利用可能になれば，大量の文書の入力に費やされる手間は軽減する．

しかし計算機による手書き文字認識は，人間に近い認識能力を獲得しているとはいい難い．今日まで認識のための様々な手法が考案され，優れた認識能力を示しているものもあるが，どの手法も一長一短があり，完全なものは得られていない．

ところで，認識手法によって認識できる文字とできない文字の集合は異なる場合が多い．そこで，複数の認識手法を統合することで，認識性能の向上を計る試みが近年なされている[2)~11)]．

本論文では，手書き文字認識の際に使用されるパターン特徴が数多く提案されていることをふまえ，複数特徴を統合する方式について考察する．

2. 複数特徴の統合

従来より複数の文字認識手法の統合について様々な試みがなされ，関連する研究が幾つか報告されている．このような統合手法の多くは，複数の認識特徴を直列に用いた階層的分類方式[8)~11)]と，複数の認識特徴を同列に評価する並列分類方式[2)~7)]に大別される．

階層的分類方式は，一般に大分類・詳細分類という

† 九州工業大学情報工学部
Faculty of Compluter Science and Systems Engineering, Kyushu Institute of Technology
☆ 現在，現在，松下電器産業株式会社
Presently with Presently with Matsushita Electoric Industrial Co., Ltd.

ように，段階的な分類を行う方式である．階層的分類方式の問題点として，設計の難しさがある．どのような順番で複数の認識特徴を評価するか，ということは認識器の設計者の直感や試行錯誤によるところが大きい．したがって，設計者が認識する対象と認識特徴の性質を熟知していないと，高い統合効果は得られないと考える．

並列分類方式は複数の認識特徴から得られるカテゴリー（モデル）ごとの信頼度を，何らかの統合関数で総合評価する方式である．並列分類方式における問題点は，計算量の多さと統合関数の統合における意味の分かりにくさである．この統合方式は，すべての認識特徴による結果を並列に統合するため，認識特徴の数に対応した計算量が必要となる．また，各認識特徴の結果を並列に統合評価することは，各特徴が統合にどのように関与しているのかつかみにくくする．

3. 前提条件

パターン識別器（識別器）の構成はいくつか考えられるが，以降では図 1 のような単純な構成の識別器の 1 つの使用を仮定する．この識別器に未知パターンを入力すると，まずパターン特徴が抽出される．識別器は学習によって作られた複数のモデルの情報を持っており，各モデルはそれぞれあるパターンの集まりを代表している．識別器は未知パターンの特徴をそれらのモデルと照らし合わせ，入力されたパターンが各モデルに属する確率（事後確率）を出力する．

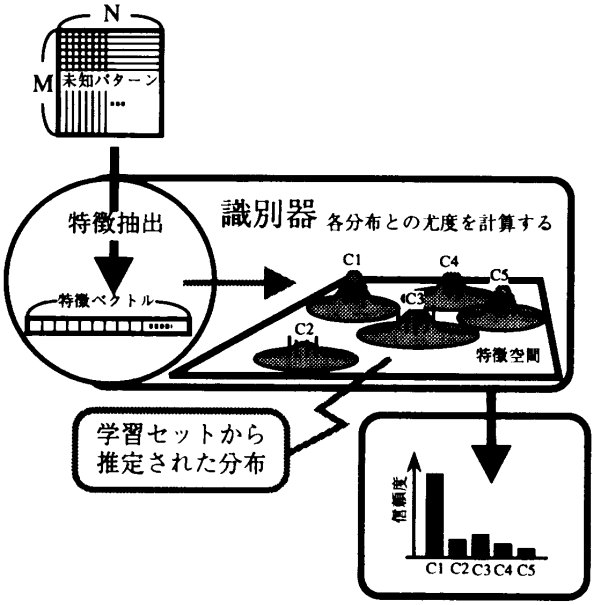

図 1 パターン識別器
Fig. 1 A pattern recognizer.

4. 既存の統合法

並列分類方式を例にとって，既存の統合法について考える．並列統合手法では，事後確率 P を次式で計算する．

$$P(C_k|X) = \sum_{i=1}^{m} w_i P(C_k|\alpha_i(X)). \quad (1)$$

$P(C_k|\alpha_i(X))$ は，未知パターン X から得られたパターン特徴 $\alpha_i(X)$ がカテゴリ C_k であるときの事後確率を表している．パターン X が入力されたときカテゴリが C_k であるための確率は，各識別器の出力の線型和で表現されている．ただし，$w_i > 0$, $\sum w_i = 1$ を満たすものとする．

本来，求める確率 $P(C_k|X)$ を正確に推定できればその方が望ましい．しかし，原パターンのままで認識を行うことはきわめて次元の高い空間での分布関数の推定を必要とするため困難である．したがって，通常はそれよりも低い次元の特徴空間に写像して分布の推定を行う．式 (1) は，次元の低いパターン特徴を用いることにより，学習時の推定誤差を少なくするとともに，いくつかの特徴量について和をとることにより，生じた誤差をさらに少なくすることを期待している．ただし，式 (1) の事後確率の計算は，ベイズルールにより式 (2) のように変形して計算する．

$$P(C_k|X) = P(C_k) \sum_{i=1}^{m} w_i \frac{P(\alpha_i(X)|C_k)}{P(\alpha_i(X))}. \quad (2)$$

荷重をどれも等しく指定するのが，均等荷重評価和という統合手法である．識別器の数を m としたときの荷重を $1/m$ とすることで，事後確率の平均値が求められる．

このほかに非線型な統合を行う図 2 のような方法がある．

多数決法は，識別器に最大評価のカテゴリに対する

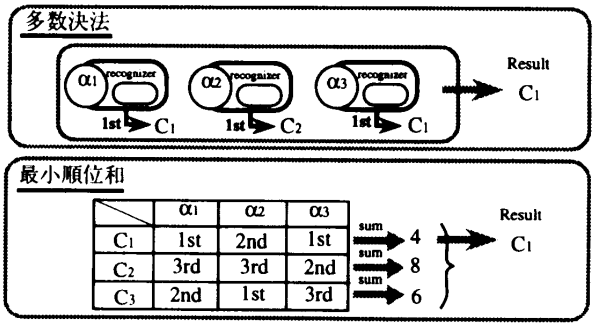

図 2 並列分類手法
Fig. 2 Other integration techniques.

投票権を与えたとき，最多得票のカテゴリを統合結果とする方法である．この方法の延長に全者一致法があり，これはすべての識別器の最大評価のカテゴリが一致したときのみ統合結果を出力する．また最小順位和は，識別器ごとにカテゴリに対して順位をつけ，順位の合計で統合結果を求める方法である[2)~7)]．

5. EID3

認識対象に対する各特徴の特性を知らなくては認識特徴の統合ができない場合，その統合には認識器を設計する際の設計者による管理指導（スーパーバイズ）が必要になる．今後増えると予想される，複数の認識手法を統合したいという要求に応えるためには，前節で紹介した並列分類方式のような教師なし（アンスーパバイズド）の統合方法が期待される．

しかし，並列分類方式には次のような問題がある．

(a) 統合する識別器の数に比例して計算量が増大する．

(b) 各識別器の出力を並列に統合するため，統合による性能向上（統合効果）がどのような理由で得られたか分かりにくい．

これらの問題点は，階層的かつ逐次的な絞り込みを行う階層的分類方式ではうまく解決できている．

そこで，識別器の階層的な構造による統合をアンスーパーバイズに設計する手段が存在すれば，効率よく統合効果を得ることができると考える．

サンプル集合をクラスタリングすることにより1つの認識特徴を用いた識別器が構築されると仮定した場合，本論文では，**図 3** のような仕組みで認識特徴を統合する方法を提案する．

この方法では，未知パターンが先頭（ルートノード）の識別器に入力されると，識別器固有の特徴抽出がなされ，識別により適切な次の識別器を選択する．終端（葉ノード）の識別器にたどりつくまで識別を行ってパスを作り，パス上の識別器の結果を統合することで，効率の良い統合を実現する．

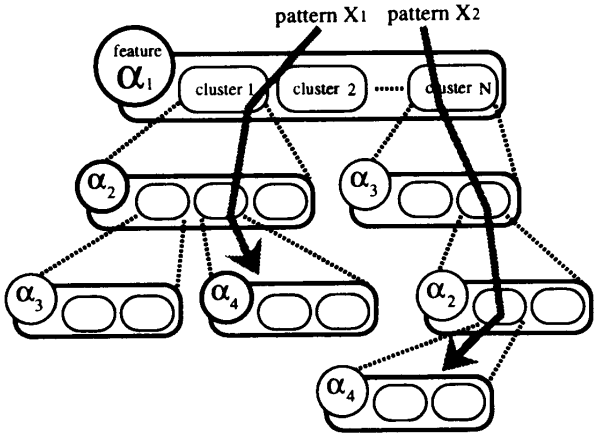

図 3 EID3
Fig. 3 A concept of EID3.

このようにして作られたツリーを，本論文では EID3 (Extended Induce Tree) とよぶ[9)~11)]．

Quinlan らの ID3[1)]は，属性値が記号で表されたデータを対象にしているのに対し，EDI3 では属性値が連続値をとる数値で表現されたデータも扱える．たとえば，ID3 では各人の背の高さと体重を表した属性値が記号のデータ $\{(High, Medium), (Low, Heavy)\}$ は扱えるが，属性値が連続値をとる数値のデータ $\{(183.9, 70), (160.1, 82.9)\}$ をそのままでは扱うことができない．一方，EID3 では学習時に決定木を構成しながら同時にクラスタリングにより，数値データの集合をクラス分けするのに適切なモデル化を行う（たとえば，*High*, *Medium* の2つのクラスからなることを決め，それぞれの平均と分散を求めて分布関数をガウス分布で近似する）．

EID3 では，パターンの属するカテゴリを特定するのに有効な特徴を漸次選択して判定を行い，その結果を統合することによる性能の向上と，パターンの階層的な絞り込みを行うことによる効率の向上（計算量を低減する）を可能にしている．

EID3 を構成するアルゴリズムとして，次のものを提案する．

(i) 学習の対象パターン集合（初回は学習サンプル全体）に対し，統合に用いるすべての認識特徴を抽出し，それぞれの特徴空間にすべての学習サンプルを記録する．

(ii) (i) で記録した各特徴空間に対し，クラスタリングを行い，クラスタリング結果に対し評価を行う．

(iii) (ii) において，最も良いクラスタリング評価を得た特徴空間を対象パターン集合における最適な特徴空間として撰択する．

(iv) (iii) で得た特徴空間内の各クラスタについて，クラスタ情報（平均，分散など）を記録する．そして，クラスタ内のサンプル集合を新たに学習の対象パターン集合として，(i) からの操作を分割されたパターン集合の中に1つのカテゴリーのサンプルのみしか含まれなくなるまで繰り返す（実際には，決定木の最大長を定め，それを越える場合はそこで打ち切っている）．

上記のアルゴリズムは，各特徴におけるクラスタリングの結果を解析して，学習の対象となるパターン集合を分割するのに適切な特徴を定めている．

全体として識別誤りを最小にする最適な特徴を決め

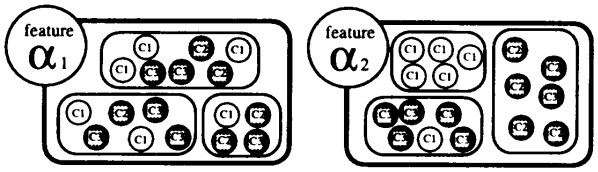

図4 異なる認識特徴におけるカテゴリ分布
Fig. 4 Distribution of categories for each feature.

ていくためには，多くの可能な組合せを考慮しなければならない．実際，最初の識別器（ルートノード）における最適な特徴は，可能なすべてのEID3を考え，その中で識別誤りが最小のEID3のルートノードで選択された特徴がそれである．

しかし，EID3として構成可能な特徴配置をすべて考慮することは，計算量の点から不可能に近い．そこで，局所情報に基づき漸次特徴を決定していく方法をとるのである．

特徴選択のための基準（局所情報）はいくつか考えることができる．ここでは，認識時の効率をあげるために式(3)を用いて評価を行う．

$$H = \sum_{i=0}^{K} \sum_{j=0}^{M} -q_i p_{ij} \log_2 P_{ij}. \qquad (3)$$

ここで，M はカテゴリ数，K はクラスタ数，p は各クラスタにおけるカテゴリの生起確率，q はクラスタの生起確率である．この式は，カテゴリ分布のエントロピを表している．

このエントロピが小さいほど，クラスタ内のカテゴリ構成は単純である．図4に示すクラスタリング結果の場合，α_1 と α_2 では，α_2 のほうがエントロピ H が小さく，このパターン集合における認識特徴として適しているとする．以上のアルゴリズムにより，効率よく統合効果をあげるEID3が構築される．

次に，EID3の識別のための評価値計算を考える．各識別器は，カテゴリの混在したクラスタの集まりで構成されており，各クラスタは，その構成要素の出現確率が正規分布に従うものとする☆．このときのEID3の出力する信頼度 P は式(4)で表される．

$$P(C_k|X) = \sum_{i=1}^{n(X)} W_{\beta_{(i)}} P\left(C_k|\alpha_{\beta_{(i)}} \in D_{\beta_{(i)}}\right). \qquad (4)$$

ここで $P(C_k|\alpha \in D)$ は，特徴ベクトル $\alpha(X)$ が領域 D に含まれるとき，未知パターン X のカテゴリが C_k である確率を示す．式(4)の右辺では，$\beta_{(i)}$ は識別過程で選択されたパス β の i 番目（段目）を表す．$D_{\beta_{(i)}}$ は i 段目でクラスタ分けの対象となる領域（i 段目の識別器でいくつかのクラスに振り分けられるが，$D_{\beta_{(i)}}$ はそれらのデータ領域）．$n(X)$ は入力パターンが X のとき，決定木の中をたどるパスの長さを表し，w は荷重を表す．荷重は，たどったパスの長さの逆数を採用した．

☆ 学習時に求めた平均と分散から，出現確率を規定する分布関数を計算する．分布関数は正規分布を仮定する．式(4)の P の計算はこの正規分布をもとに計算する．

6. 実　　験

6.1 実験環境

実験には，通産省電子総合研究所から提供されている手書き文字データベース：ETL6から手書き数字（0～9）の部分を使用した．元データには輝度の傾斜やノイズがあるため，以下の処理を施した．

- 平滑化フィルタによるノイズ除去
- ロバートフィルタによる輪郭線抽出
- 判別基準法による二値化[7]
- 縦横比を保存した外接矩形による正規化

こうして得られた10カテゴリのデータ，全13000サンプルのうち，学習用として奇数シリアル番号の6500個，テスト用として，偶数シリアル番号の6500個を用いた．

統合すべき認識特徴は，以下のものを使用した[8]．

- メッシュ特徴（36次元，64次元）
- 周辺分布特徴（24次元，32次元）
- ストローク密度特徴（16次元，32次元）
- 方向線素特徴（36次元，64次元）

また各識別器は，学習用のパターンから得られる認識特徴の集合を，変数選択の後，K 平均法によりクラスタリングし，得られたクラスタを多次元正規分布と仮定して用いる．ここで，予備実験を行い（誤認識率 × 10）＋認識拒絶率 をコスト評価関数し，これが最小となるように選んだパラメータを用いて比較を行った．

6.2 正解率による検証

比較のために，単特徴を単純に適用した場合の正解率を実験により求めた．その結果を図5に示す．

上記の識別器は学習用サンプルにおいて，各カテゴリごとに1クラスタ（全10クラスタ）を割り当て，学習したものである．最も正解率の高かったのは，64次元の方向線素特徴を用いた識別器で，97.0%であった．

次に，先に述べた3つの並列分類手法による上記特徴の統合結果を図6に示す．

結果として，実験に用いたどの統合法も，単一特徴の識別器よりも高い正解率を示した．最も統合効果が

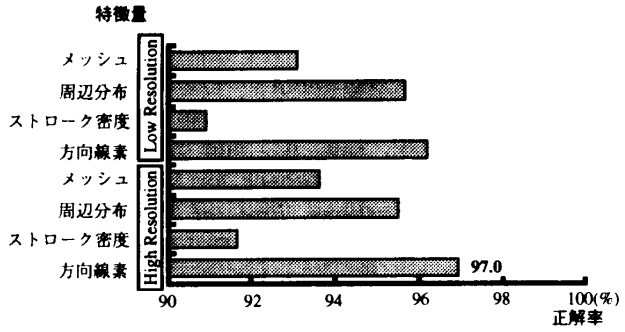

図5 単特徴の識別器の正解率

Fig. 5 Performance on correct identification by single feature.

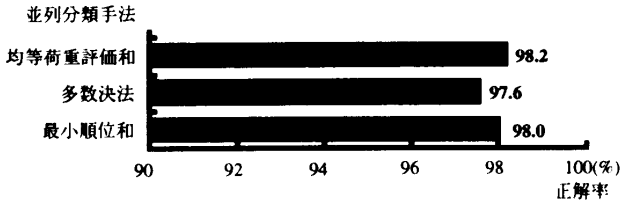

図6 8特徴の並列分類手法による正解率

Fig. 6 Performance on correct identification by integrated 8 featurs.

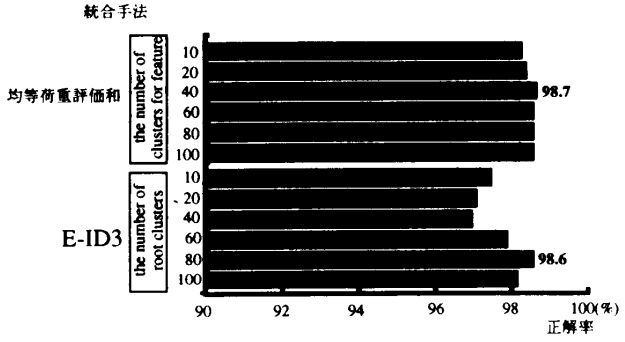

図7 クラスタリング条件を変えた正解率

Fig. 7 Performance on correct identification for each number of cluster.

現れたのは，均等荷重評価和の 98.2% であった．

次に，並列分類方式の中から安定して高い正解率を示した均等荷重評価和を選び，EID3 との比較実験を，クラスタリング条件を変えて行う．並列分類方式に用いる各識別器のカテゴリごとのクラスタリングの際のクラスタ数を 1・2・4・6・8・10 （全クラスタ数はそれぞれ 10・20・40・60・80・100）と変えた．EID3 は，各ノードにおいて対象とするサンプル集合のカテゴリの種類をクラスタ数としてクラスタリングを施したが，ルートノードの識別器についてはクラスタ数を 10・20・40・60・80・100 と変えて正解率を求めた．この実験結果を図 7 に示す．

この結果，均等荷重評価和の統合による最高の正解率が 98.7%，EID3 による最高の正解率が 98.6% と両者はほぼ同様の性能を示した．しかし図から明らかなように，クラスタリング条件を変えた場合の性能は，均等荷重評価和の方が安定しており，EID3 ではルートノードのクラスタ数を変えると識別木の構成が変わるため認識率が多少変動している．

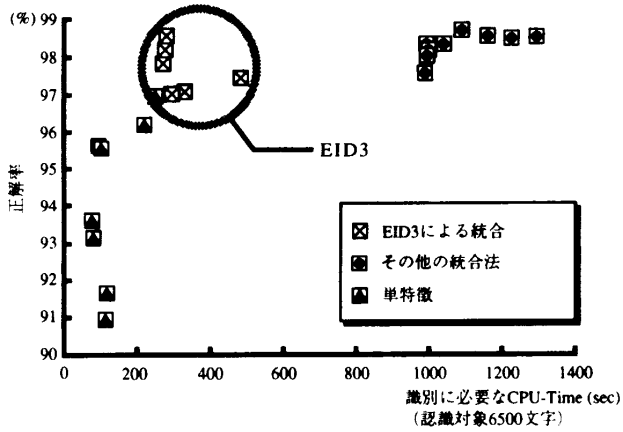

図8 識別時間・正解率グラフ

Fig. 8 Chart of CPU-Time and performance on identification.

この実験において，識別にかかった時間と正解率との関係をプロットしたグラフを図 8 に示す．▲は単特徴の認識器の認識結果，×は EID3 でルートノードのクラスタ数を変えて構成した識別器の認識結果，●は並列的な統合方法を用いた識別器の認識結果をそれぞれ表している．

EID3 は，少ない識別時間で，統合効果により高い正解率を得ている．識別時間が少ない理由は 2 つある．1 つは，識別の大部分が 2〜3 個の識別器の統合で結果を出している（全体的に木の段数が浅い）ためである☆．もう 1 つは，木構造に配置された識別器の中でも，葉に近いものはそれ以前の識別器の絞り込みにより，マッチング回数が少ないためである．

6.3 追加実験

EID3 が短い時間で統合効果を得ることに着目して，EID3 による統合結果を 1 つの識別器の出力として見なし，既存の統合法との融合を考える．

図 9 は，4 種類の単特徴の識別器の正解率とともに，いくつかの組合せで統合したときの正解率である．

実験結果から，4 特徴の EID3 による統合，4 つの単一特徴による識別器の均等荷重評価和による統合よりも，「4 特徴の EID3 + 4 つの単一特徴による識別器」を均等荷重評価和で統合したものが高い正解率

☆ ルートノードの識別器の全クラスタ数が 10 であるときの EID3 の構成は，識別器が全 54 個，木の深さが最大で 5 段であった．

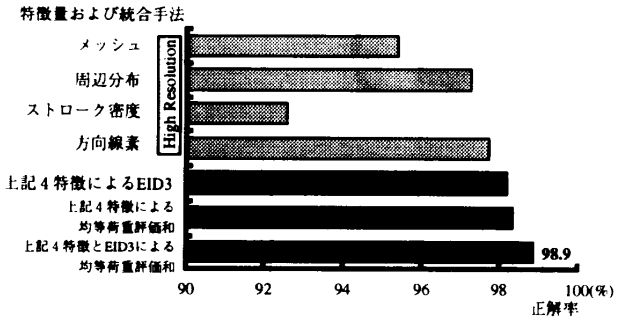

図9 EID3を含む並列分類手法による統合
Fig. 9 Performance on correct identification by integrated EID3 and some features.

(98.9%) を示していることが分かる．このことから，EID3は局所的な視点ではあるが学習パターンを分離するために適した特徴を選択することにより，新しい特徴の組合せを選び出していると考えられる．

7. むすび

実験結果から，各種の統合手法による複数特徴の統合は，正解率の向上に貢献することが確認できた．

提案したEID3は並列分類手法に近い正解率を得た．また，識別時間に関しては，統合する認識特徴の数にかかわらず，単一特徴の識別器の2～3倍程度の識別時間で，並列分類手法と同程度の正解率が得られることが確認できた．

さらに，EID3を1つの識別器として見なし，統合する対象に加えることで，既存の統合による統合効果をより向上できることが確認できた．

今後，新たな特徴選択基準の開発やクラスタリング評価の際に先読みなどを行うことにより，より高い統合効果を効率よく得るEID3を構築することができると考える．

謝辞 本研究を行うにあたって，貴重な手書き文字データを提供していただいた通産省工業技術院電子技術総合研究所の諸氏に感謝いたします．

参考文献

1) Quinlan, J.R.: Induction of Decision Trees, *Machine Learning*, Vol.1, Kluwer Academic Publishers, pp.81–106 (1986).
2) 松井，山下，若原，吉室：文字認識アルゴリズムの複合化手法の検討—第1回文字認識技術コンテストの結果より—，信学技報，PRU92-33, pp.65–72 (1992).
3) Noumi, T., Matsui, T., Yamashita, I., Wakahara, T., and Tsutsumida, T.: Result of Second IPTP Character Recognition Competition and Studies on Multi-Expert Handwritten Numeral Recognition, *Proc. of The Fourth International Workshop on Frontiers in Handwritten Recognition*, pp.338–346 (1994).
4) Huang, Y.S., Liu, K., and Suen, C.Y.: A Neural Network Approach for Multi-Classifier Recognition Systems, *Proc. of The Fourth International Workshop on Frontiers in Handwritten Recognition*, pp.235–244 (1994).
5) Huang, Y.S. and Suen, S.Y.: Combination of Multiple Classifiers with Measurement values, *Proc. of 2nd International Conference on Document Analysis and Recognition*, pp.598–601 (1993).
6) 梅田三千雄：マルチフォント印刷漢字の分類，信学論 (D)，Vol.J62-D, No.2, pp.133–140 (1979).
7) 萩田，梅田，増田：3つの概形特徴を用いた手書き漢字の分類，信学論 (D)，Vol.J63-D, No.12, pp.1096–1102 (1980).
8) 坂井，平井，河田，天野，森：印刷漢字OCRのためのシミュレーションシステム，情報処理，vol.17, no.7, pp.4–46 (1976).
9) 井藤，大橋，江島：複数特徴による階層的識別器の自動設計に関する一手法，画像の認識理解シンポジウム (MIRU '94) 講演文集，I, pp.137–144 (1994).
10) 井藤，大橋，江島：複数特徴を考慮した識別器E-ID3の統合方法は？，*The 17th Symposium on Information Theory and Its Applications*, pp.779–782 (1994).
11) Ito, Y., Ohashi, T., and Ejima, T.: Considerations on Designing a Decision-Tree with Multiple Features, *Proc. of The Fourth International Workshop on Frontiers in Handwritten Recognition*, pp.362–369 (1994).
12) 舟久保登：パターン認識，pp.128–130, 共立出版 (1991).
13) 橋本新一郎：文字認識概論，pp.57–139, 電子通信協会 (1982).

(平成7年3月17日受付)
(平成8年2月 7 日採録)

井藤 好克 (正会員)
昭和44年生．平成5年九州工業大学情報工学部知能情報工学科卒業．平成7年同大大学院情報工学研究科情報科学専攻博士前期課程修了．同年松下電器産業（株）入社．パターン認識処理の研究・開発に従事．修士（情報工学）．

大橋 健（正会員）

昭和 42 年生．平成 3 年長岡技術科学大学大学院工学研究科修士課程修了．同年九州工業大学情報工学部知能情報工学科助手．パターン認識とこれを活用したヒューマン・インタフェイスなどの研究に従事．工学修士．電子情報通信学会，ソフトウェア科学会，IEEE 各会員．

江島 俊朗（正会員）

1951 年生．1978 年東北大学大学院博士課程修了．現在，九州工業大学情報工学部知能情報工学科教授．文字認識，ニューラルネットワーク，画像処理，ヒューマンコンピュータインタラクション，自律分散型ロボットなどの研究に従事．工学博士．IEEE，電子情報通信学会，神経回路学会各会員．