

複数特徴整合を用いた重ね書き文字の分離抽出

松尾 賢一・堀 元*

Separation and Extraction of Overwritten Characters Using Plural Features

Ken'ichi MATSUO, Hajime HORI

In this paper, we propose a method for separating for overwritten characters on examination paper. The overwritten marks on examination paper are the target characters of separation and extraction in this study. A mark which is a character pattern is divided as a stroke between the feature points. The temporary character group which combined the stroke mutually is created using this divided stroke. It is determined by calculated similarity whether temporary character in the group is a mark. Validity is calculated from the average value of the similarity in a temporary character group. The temporary character group which has justification most is chosen as the result of the separated character. An experiment has been performed to examine the performance of the proposed method. As a result of the experiment, separated extraction rate from the overwritten character of the 200 characters data was 83%.

1. はじめに

文字認識手法の多くは、あらかじめ1文字単位で切り出された文字パターンを処理対象としている^{[1][2][3]}。一般的に文字認識では、1文字ごとに切り出した入力パターンに対して、辞書パターン内から最も類似性の高いパターンを出力結果とする^[4]。このとき、文字同士が重畳した入力パターンを認識器に与えても、辞書パターンに存在する何らかの文字パターンを認識結果として出力する。

つまり、現在の文字認識技術は、1文字ごとに切り出されていること、あるいは、辞書パターンに登録されている文字種だけを認識器に入力される前提において成り立っている。したがって、文字同士が重畳しているパターンや、辞書に存在しないパターンに対する認識器のロバスト性は、極めて低いといえる。

これらの問題に対して、文字が重畳している入力パターンでは、何らかの方法で重畳をなくして文字同士を1文字単位に分離して認識する処理、あるいは、重畳パターン領域内を大きさや傾きを網羅的に変化させた辞書パターンによってスキャンし、最大の類似性が得られる領域を認識結果とする処理のいずれかが必要である。この文字が重畳している一例として、重ね書き文字があげ

られる。特に、同色で重ね書きされた文字は、1文字単位の切り出しが非常に困難である。

この重ね書き文字に対する切り出しや分離の手法として、入力画像の書式における前知識を用いて枠線と交差した文字領域を分離する手法^[5]、色差を用いて文字や枠線を分離抽出する手法^{[6][7][8]}、重ね書き文字が連結した閉領域と標準パターンとのずらし類似度によって1文字単位の文字の位置を検出する手法^[9]が提案されている。

しかしながら、[5-8]の手法では、同色で重ね書きされた自由手書き文字を分離することができない。また、手法^[9]では、同色で重ね書きされた自由手書き文字に適用できるが、切り出したパターン内に他のパターンのストロークがそのまま切り出されてしまう問題があった。

本論文は、同色で重ね書きされた自由手書き文字の分離抽出を実現するために、テスト答案に重ね書きされた採点記号や数字を対象とした分離抽出手法の提案と、分離抽出実験による本手法の有効性について述べる。

2. 重畳採点記号の分離抽出について

2.1. 文字の定義

採点者がテスト答案に記入した筆記の全てを採点領域とするとき、この採点領域には、「○」、「△」、「×」、

* 奈良先端科学技術大学院大学 情報科学研究科

「チェック」の4種類の採点記号と「0」から「9」までの数字が、単一あるいはお互いに接触、重畳した状態で混在している。これらの採点記号および数字が接触、重畳している文字を「重畳文字」、重畳していない文字を「単一文字」、採点記号同士が重なっている採点記号を「重畳採点記号」と定義する。

2.2. 提案手法の流れ

採点領域において同色で重ね書きされた採点記号を分離抽出する提案手法の流れを図1に示す。

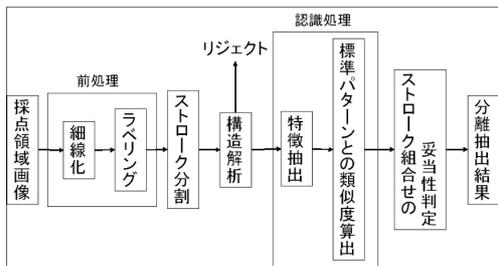


図1 提案手法の流れ

提案手法では、重畳採点記号をストローク単位に分解し、得られたストロークを網羅的に組合せる。そして、ストロークを組合せたパターンに構造解析を施し、採点記号の構造的特徴をもたないパターンをリジェクトすることで計算量の削減を図る。リジェクトされなかったパターンに認識処理を行い、求めた類似度からストローク組合せの妥当度を判定する。最後に、ストローク組合せの妥当度に基づき、採点記号として最も妥当な形状をもつ部分ストロークの組合せを一意に決定する。

2.3. 前処理

図2(a)の答案画像から [6] の手法を用いて、同図(b)に示す採点領域を「1」、背景を「0」の画素値とする2値画像を作成する。この2値画像を「採点領域画像」とする。

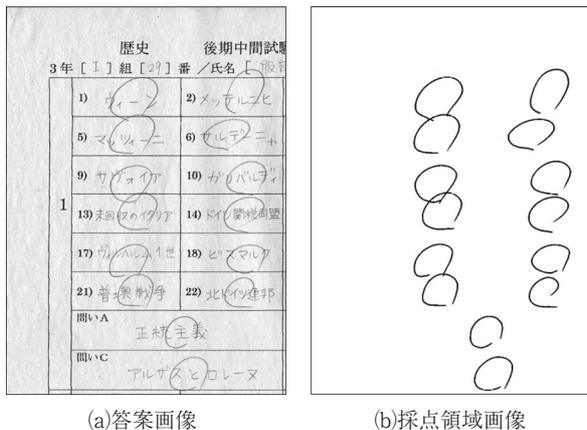


図2 採点領域画像の一例

2.3.1. ストローク分解による仮パターン群列の生成

図3にストローク分解による仮パターン群列の生成過程を示す。同図(a)の入力パターンに対して、Hilditchの方法^[10]によって細線パターンを得る。この細線パターンに対して、3つの特徴点である「端点」、「交差点」、「分岐点」を検出する。次に、同図(b)のように3つの特徴点間で、ストロークを分解する。このストロークを「部分ストローク」と定義する。ここで、N本の部分ストロークに分解されたとき、各部分ストロークに対し、ラベルL(L=1,2,...,N)を与える。このN本の部分ストロークを網羅的に組合せ、図3(c)に示す仮パターン群列を生成する。部分ストローク組合せの総数をx(x=1,...,N)とすると、この仮パターン群xは、N本のストロークから、x本のストロークを取り出して生成される全ての仮パターンの集合となる。

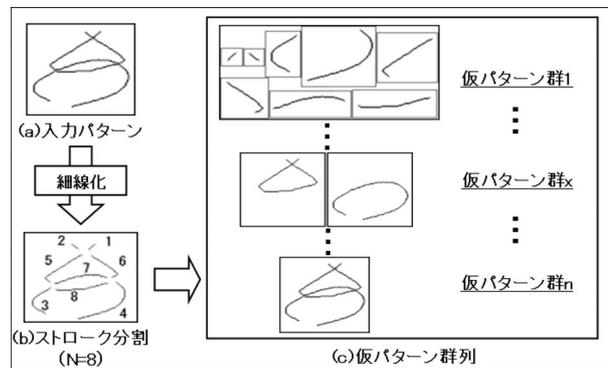


図3 仮パターン群列の生成過程

2.4. 構造解析

ここで、各仮パターン群の全てのパターンに対して、端点、交差点、分岐点における交差数と各特徴点の位置関係によって採点記号と同じ構造特徴をもつパターンを検索する。

各特徴点における交差数CN(crossing number)は、

$$CN = \sum_{k=1}^8 x((k+1) \bmod 8) - x(k) \quad (1)$$

によって求められる^[11]。

この交差数CNは、図4(a)の注目画素x(0)の8近傍画素x(1)からx(8)を環状に1周走査したときに、各画素の値が0から1、もしくは、1から0に変化する回数である。この走査を仮パターンの全てに対して行い、交差数CNによって図4(b)に示す8近傍パターンとなる特徴点(端点、分岐点、交差点)を仮パターン内で検出する。

次に、検出した特徴点の位置関係から、採点記号を同じ構造をもつ仮パターンだけ検出する。ここで、4つの

採点記号の構造特徴を、

採点記号「○」の構造特徴

1. 端点は2つ以下
2. 端点は上部4割に存在しない
3. 交差点(分岐点)は, 上部5割に存在しない

採点記号「△」の構造特徴

1. 端点は2つ以下
2. 端点は下部4割に存在しない
3. 交差点(分岐点)は下部4割に存在しない

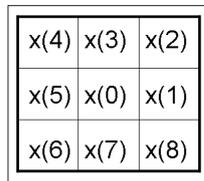
採点記号「×」の構造特徴

1. 端点は4つ
2. 端点は上から4~6割, 左から4~6割に存在しない
3. 上から5割, 左から5割に1つ, 上から5割, 右から5割に1つ, 下から5割, 左から5割に1つ, 下から5割, 右から5割に1つ

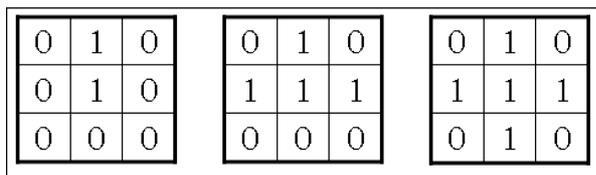
採点記号「チェックマーク」の構造特徴

1. 端点は2つ
2. 端点は上から1割, 右から1割に1つ, 左から3割に1つ
3. 交差点(分岐点)は存在しない

と定義し, 各々を採点記号の構造特徴条件とする。



(a) 8近傍



端点 分岐点 交差点

(b)特徴点

図4 注目画素の8近傍

4つの採点記号の構造特徴の条件を全て満たす仮パターンを含んだ仮パターン群は, この段階で仮パターン群列からリジェクトする。複数の採点記号と構造特徴が合致する仮パターンについては, 後続の処理によって採点記号が決定される。

2.5. 特徴抽出とパターン認識処理

特徴抽出は, パターン特有の特徴だけを抽出する処理である。提案手法では, メッシュ特徴, ペリフェラル特徴, ストローク密度分布特徴の3つの特徴量を採点記号

と仮パターンから抽出する。

2.5.1. メッシュ特徴

メッシュ特徴^[13]は, 対象パターンを横 m ×縦 n 画素の大きさのメッシュに区切ったときの, メッシュ面積 $m \times n$ に対する黒画素数の割合である。提案手法では, 仮パターンの外接矩形を 16×16 のメッシュに区切り, 特徴を抽出する。

2.5.2. ペリフェラル特徴

対象パターンを外枠の1辺につき8分割し, 一方の外枠から対面する外枠に向かって走査する。そして, 最初の文字ストロークに出会うまでの背景領域の面積を求める。同様の処理を上下左右4つの外枠について行う。ここで得られる値は1次ペリフェラル特徴^[13]と呼ばれる。さらに, 1次ペリフェラル特徴と同様にして2番目の文字ストロークに出会うまでの背景領域の面積は2次ペリフェラル特徴となる。提案手法では, 仮パターンから1次と2次両方のペリフェラル特徴を抽出する。

2.5.3. ストローク密度分布特徴

横 x ×縦 y 画素の対象パターン中の着目画素 $P(x,y)$ から上下左右方向に画素を走査したとき, 画素値が0から1に変化する回数を求める。このとき, 対象パターンを縦横4分割したメッシュ領域内で求めた変化回数を平均化した特徴がストローク密度分布特徴^[14]となる。

2.5.4. 複合特徴の生成

前述した3つの特徴量を統合した複合特徴ベクトルを生成する。抽出した各特徴量の次元数はメッシュ特徴が256次元, ペリフェラル特徴が64次元, そして, ストローク密度分布特徴が64次元である。したがって, 統合された複合特徴の次元数は, 3つの次元数を総和した384次元となる。

2.5.5. 類似度の算出

標準パターンは, 入力パターンを認識するために必要な文字カテゴリごとに分けられたパターンの総称である。この標準パターンの作成に用いられたパターンを学習パターン, 用いられていないパターンを未知パターンと呼ぶ。

この標準パターンと入力パターンの両方に対して, 前述した特徴ベクトルを得る。そして, この特徴ベクトル間で類似度を求める。

ここで, 仮パターンと標準パターンが d 次元の特徴ベクトルで表されるとき, 仮パターン特徴ベクトル $f = (f_1, f_2, \dots, f_d)$ と, 標準パターン特徴ベクトル $g = (g_1, g_2, \dots, g_d)$ との内積は

$$\langle f, g \rangle = \sum_{k=1}^d (f_k * g_k) \tag{2}$$

となる。このとき、特徴ベクトル f, g のノルム $\|f\|, \|g\|$ が

$$\|f\| = \sqrt{\sum_{k=1}^d |f_k|^2} \tag{3}$$

$$\|g\| = \sqrt{\sum_{k=1}^d |g_k|^2} \tag{4}$$

であるとき、類似度 r は、

$$r = \frac{\langle f, g \rangle}{\|f\| * \|g\|} \tag{5}$$

で表される。

この類似度 r の値域は、0から1であり、1に近ければ特徴ベクトル f, g 間の類似性が高いことを表している。

ある仮パターンに対して、構造特徴の条件を満たす採点記号の標準パターンとの類似度を算出する。ここで、算出した類似度は、後のストローク組合せで生成された仮パターン群中の仮パターンの妥当度の算出に用いる。

また、構造解析において仮パターンが複数の採点記号の構造特徴の条件を満たすときには、条件を満たした全ての採点記号の標準パターン間で類似度を算出する。このとき、類似度が最大となる採点記号のパターンに仮パターンを分類する。

2.6. 妥当性判定による仮パターン群の選択

重畳採点記号が交差前の単一文字に分離されるストロークの組合せは、作成した仮パターン群の中の唯一である。ここで、重畳採点記号が単一文字に分離される唯一のストローク組み合わせである仮パターン群を「正解仮パターン群」、それ以外の群を「不正解仮パターン群」と定義する。図4に妥当性の判定による仮パターン群選択の概念図を示す。

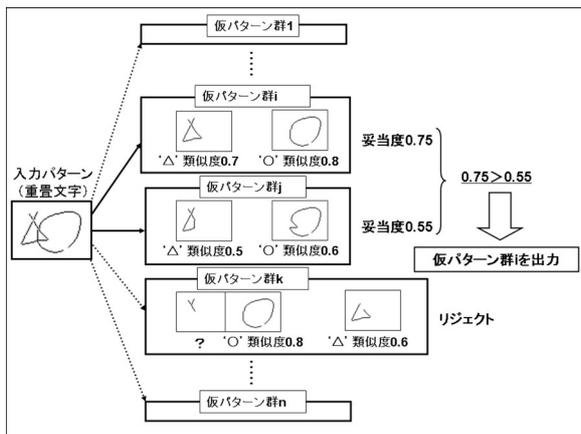


図4 妥当度判定による仮パターン群の選択

図4の仮パターン群 k は、構造解析の処理においてどの採点記号とも構造特徴条件を満たさないパターンを含んでいるためにリジェクトされる。一方、仮パターン群 i と j は、採点記号の構造特徴の条件を満たした仮パターン群である。よって、ストローク組合せの妥当度が算出される。この妥当度は、仮パターン群内の仮パターンの類似度 r の平均値とする。図中では、仮パターン群 i, j の妥当度は、パターン群内の類似度 r の平均値により、各々0.75, 0.55となっている。この i と j の妥当度を比較した結果、妥当度が最も高い仮パターン群 i が重畳した採点記号を分離した結果として選択され、分離抽出結果として出力される。

3. 実験及び考察

この章では、提案手法の有効性を重畳採点記号に対する分離抽出実験によって明らかにするとともに、結果に対する考察について述べる。

3.1. 実験条件

3.1.1. パターンの作成

5人の筆者から「○」、「△」、「×」、「チェック」の4種の単一採点記号を各々100パターン、計400パターンを収集した。収集したパターンに対して、メッシュ特徴、ペリフェラル特徴、ストローク密度分布特徴、そして、3つの特徴を総和した複合特徴を別々に文字パターン単位で抽出する。そして、各々の特徴ベクトルを文字カテゴリで平均して標準パターンを作成する。また、2文字の採点記号が重なるように3人に筆記させた200文字の重畳採点記号を未知パターンとする。

ここで、未知パターンにおける採点記号の種類別比率は「○」が3割、「△」が3割、「×」が2割、「チェック」が2割程度である。なお、未知パターンの筆者3人のうち1人は標準パターンの筆者と重複している。この未知パターンに対しても、標準パターン同様、4種の特徴ベクトルを抽出する。

3.1.2. 分離抽出の正誤の定義

分離抽出の結果を、目視によって分類する。図5において、入力パターンである重畳採点記号が同図内(a)のように、単一文字の状態でも正確に採点記号が出力された仮パターン群であるとき「分離抽出」、同図内(b)のように不正確な採点記号をもつ仮パターン群が出力されたとき、「分離誤り」とする。

このとき、構造解析でリジェクトされた仮パターン群については、同図内(c)のように、「構造解析誤り」とす

る。ここで、分割後のストローク数が10個以上の未知パターンについても、分離抽出処理を打ち切り、「構造解析誤り」に分類する。

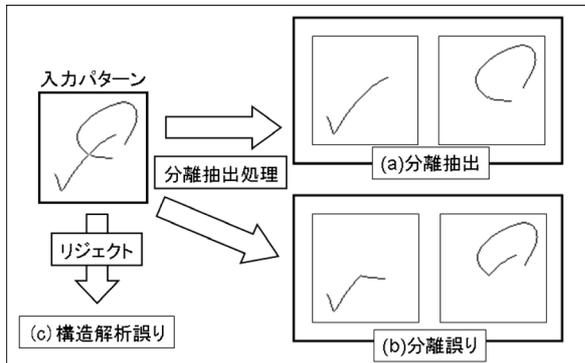


図5 分離抽出処理による仮パターン群の分類結果例

3.2. 分離抽出実験結果

未知パターン200文字の重畳採点記号を用いて、分離抽出実験を行った。4つの特徴別の分離抽出文字数および分離抽出率を表1に示す。

表1 分離抽出結果（重畳採点記号200文字当り）

特徴抽出法	分離抽出文字数	分離抽出率
メッシュ特徴	136文字	68%
ペリフェラル特徴	140文字	70%
ストローク密度分布特徴	142文字	71%
複合特徴	166文字	83%

実験の結果、提案手法によって同色の重畳採点記号を83%の精度で分離抽出できた。この複合特徴においては、3つの特徴を単一で用いたときの分離抽出率を上回る結果が得られた。

3.3. 実験結果に対する考察

3.1.1. 分離抽出失敗の内訳

重畳採点記号200文字に対する分離抽出の内訳を表2に示す。

表2 分離誤抽出の内訳

特徴抽出法	分離誤り	構造解析誤り
メッシュ特徴	52文字	12文字
ペリフェラル特徴	48文字	12文字
ストローク密度分布特徴	46文字	12文字
複合特徴	22文字	12文字

図6で分離誤りとなった同図(a)の入力パターンと、(b)の分離抽出結果を比較すると、不正解仮パターン群中の仮パターンの多くは、図6(b)の左の仮パターンのようにストロークに局所的な歪みが数多く見られた。実際、同図(c)の正解仮パターン群を比較すると、一部を除き、ほぼパターンの形状が一致している。

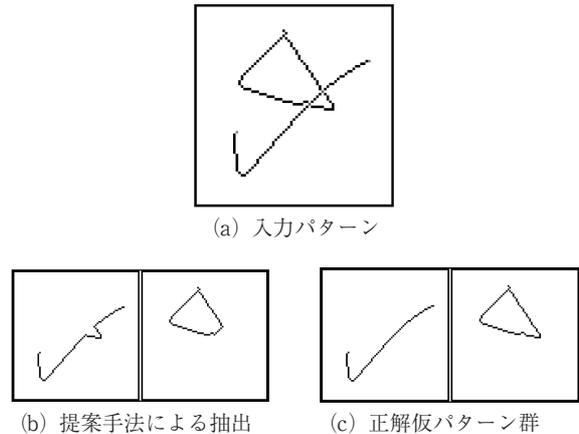


図6 認識誤り例

つまり、(a)の入力パターンに対する(b)と(c)との類似度は、共に高い値を示すことが予想される。この提案手法での類似度を用いた妥当度の算出は、パターンマッチングを基本としているため、非常に類似したパターンからの正解パターンの選択が困難であることが、分離誤りの大きな原因であったといえる。正解仮パターン群中の仮パターンであるにも関わらず、構造解析でどの採点記号の構造特徴とも合致せず、リジェクトされた図7のようなパターンが存在した。図7の入力パターン「×」は、本来「×」の構造特徴を満たさなければならない。しかしながら、入力パターンは、2.4の「×」の構造特徴の2と3の条件を満たしていない。その結果、この仮パターンをもつ全ての仮パターン群がリジェクトされるため、抽出結果として正解仮パターン群が出力されなくなる。

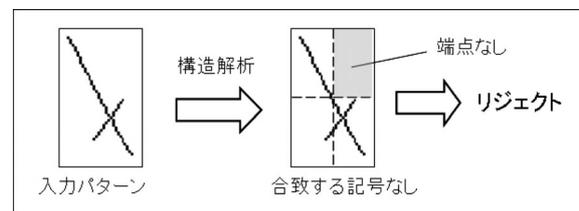


図7 構造解析でリジェクトされた正解パターン例

これに対して、構造特徴の判定条件を緩和すれば構造解析誤りを低減できるが、構造解析でリジェクトされていた不正解仮パターン群が正解パターン群となり、妥当

性の計算量や分離誤りの増加を招くことにつながる。そこで、3つの特徴点間で分解したストロークだけでなく、判別に有用なより局所的な構造特徴を抽出する方法の導入や、仮パターンそのもの形状を正規化して類似性を算出する処理の導入等が望ましい。

また、図8のように分解後のストローク数の多いパターンでは、網羅的なストローク組合せ数が膨大になり、実時間処理が困難となる。提案手法での重畳採点記号を分離するための処理時間は、分解後のストローク数に依存する。

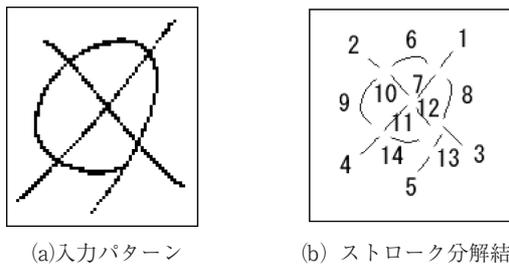


図8 ストローク数の多いパターン例

この解決として、図8のような重畳採点記号を分離するには、分解されたストロークが、元の文字ストロークを細分化した一部であるとき、他の細分化されたストローク同士で一本のストロークにして、仮パターン群を生成することが考えられる。

4. おわりに

本論文では、重畳採点記号を3つの特徴点間で分解して網羅的に組み合わせるときに得られる部分ストロークに対して、単一の採点記号間との類似度の平均である妥当度によって、重畳採点記号を単一文字同士に分離抽出する手法を提案した。また、検証実験によって提案手法に対する有効性を明らかにし、結果に対する考察について述べた。

提案手法によって、同色の重畳採点記号200文字に対して、約83%の分離抽出率を得ることができた。また、単一でなく3つの特徴抽出法を併用することで分離抽出率が向上することを明らかにした。

今後の課題として、局所的なストロークの構造の違いを判別できるように、分離抽出処理を改良するとともに、構造解析における構造特徴や分類処理の見直し、ストローク数の多い重畳採点記号への対応、さらに、分離抽出の対象を採点記号から、採点数字に拡張することが必要である。

参考文献

- [1] 中野：“文字認識・文書理解の最新動向[IV・完]”，電子情報通信学会誌, Vol.83 No7, pp.576-580,2000
- [2] 梅田：“手書き漢字認識研究の歩み”，電子情報通信学会技術報告, PRMU95-120,pp.103-110,(1995)
- [3] 後藤ら：“様々な画像に適用できる文字パターン抽出手法について～サーベイおよび構成例～”，電子情報通信学会技術報告, PRMU99-234, pp.23-30,(2000)
- [4] 中野：“文字認識・文書理解の最新動向[I]”，電子情報通信学会誌 Vol.83 No1, pp.64-68,(2000)
- [5] Alessandro L. Loerich, Luan Ling Lee: “Automatic Extraction of Filled-in Information from Bank Checks Based on Prior Knowledge About Layout Structure”, Lecture Notes In Computer Science Vol.1339 pp.322-333,(1997)
- [6] 松尾ら：“重ね書きされた文字情報の分離抽出”，2003電子情報通信学会総合大会講演論文集D12-20, p.181,(2003)
- [7] 松尾ら：“配送伝票画像からの重ね書き文字領域の分離抽出”，電学論, Vol.118-C, pp.1730-1738,(1998)
- [8] 前川ら：“銀行小切手からの記入文字と印影の分離抽出”，2004年電子情報通信学会総合大会, D-12-37,(2004)
- [9] 武井ら：“重ね書きされた文字情報の分離抽出II”，電子情報通信学会関西支部学生会 第9回学生会研究発表講演会 講演論文集, D3-1,(2004)
- [10] C.J.Hilditch：“Linear Skeletones From Square Cupboards”,(B.Melter and D.Michie ed.) Machine Intelligence 4, Eginburgh Univ. Press, p.403,(1969)
- [11] 南ら：“画像工学 -画像のエレクトロニクス-”，コロナ社, pp.102-103,(1990)
- [12] 鳥脇：“パターン認識と画像処理”，朝倉書店, pp.30-34,pp.57-59,(1992)
- [13] 萩田ら：“三つの概形特徴を用いた手書き漢字の分類”，電子通信学会論文誌,Vol.J63-D,No.12, pp.1096-1102 ,(1980)
- [14] 内藤ら：“手書き漢字認識のためのストローク密度特徴”，電子情報通信学会論文誌,Vol.J64-D, No.8,pp.757-764,(1981)