修 士 論 文

数式 OCR のための
接触記号分離に関する研究

平成 28 年度修了
三重大学大学院工学研究科
博士前期課程 情報工学専攻

川口 竜平

はじめに

近年，数式が含まれた科学技術文献の電子化を行う “InftyReader ” や，数式にスマートフォンのカメラを向けるだけでその数式の計算過程や解答を導出する “PhotoMath ” のような，数式 OCR を利用したアプリケーションがリリースされている．しかし，このような OCR は隣接する文字・記号が接触している場合に対応していない．接触記号の問題は，一般的なテキスト用の OCR でも問題となるが，数式中の文字・記号の接触は文字・記号のサイズや接触の方向（左右，上下，斜め）が多様であるため，より困難な課題である．数式の接触記号分離に関する先行研究として，画像マッチングによる分離手法 [7]，垂直線，水平線，± 45 ° の 4 方向の直線によって切断する手法 [10]，接触記号の輪郭上に検出されるコーナー特徴点の間を結ぶ線で切り分ける手法 [11] などが提案されている．しかし，画像マッチングによる手法は精度が低く，また直線によって分離する手法では数式における多様な接触状況に対応できない．本研究では，二つの記号が接触している数式中の接触記号に対して，そのストロークの芯線上の画素を用いた分離手法を提案する．

提案手法は，分離候補の生成と文字認識の評価値（文字らしさ）に基づいた分離の決定からなる．分離候補の生成には，まず接触記号から抽出したストローク芯線を二分割できる芯線上の全ての画素で，芯線を分割する．分割した芯線のそれぞれを，元の接触記号と距離変換で求めた距離値を用いて太さのある連結成分に復元することで，分離記号候補を生成する．生成した分離記号候補の全てに対して文字認識を行い，二つの記号の評価値の和が最も小さくなるものを最終的な分離結果として出力する．提案手法では，記号の接触で生じた連結成分を直接固定方向の線分で分離せずに，その連結成分の芯線上の画素に注目して分離を行っている．そのため，水平方向以外の接触に対しても，分離に適した分割線の角度を考慮せず，同様の処理で分離することができる．

科学文書用 OCR の研究のために公開されたデータベース “InftyCDB-1 ” に収蔵されている数式中の 2 つの文字・記号が接触してできた接触記号 627 個を用いて，提案手法の性能評価実験を行った．評価では，入力された接触記号を提案手法によって分離し，両方の記号を正しい記号に識別できれば分離成功とする．実験の結果，提案手法は，93.1% (584/627 個) の接触記号を正しく分離することができた．分離失敗は，分離候補の連結成

分の形状が他の字種の形状と類似する場合や，文字・記号の一部に欠損や滲み，ノイズがある場合などで，文字認識が困難な場合に多い．実際に文字・記号の形はフォントによって多様であり，フォントを推定する機能の付加によって改善可能である．また，2 つの記号の接触箇所が複数ある場合は，芯線上の一点で接触記号を分離することができないため，複数の画素を用いて分離する必要がある．

本研究では，提案する芯線分離と高精度の文字認識手法を組み合わせることにより，多様な方向に接触する数式の接触記号に対して高い精度で分離可能であることを示した．

目次

第 1 章

序論

1.1 研究の背景

1.1.1 数式の OCR

OCR(Optical Character Recognition) とは，カメラやスキャナによって画像化された文書中の文字を切り出して認識し，コンピュータで編集可能なデータに変換する技術である．近年は通常のテキスト用だけではなく，数式のための OCR が研究・開発されている．数式のための OCR は，数式中の文字・記号を切り出して認識し，また数式の構造も解析する技術である．数式 OCR の活用例として以下のようなものが挙げられる．

1. InftyReader[1]
   InftyReader は数式を含む文書のスキャン画像を LaTeX や Word，MathML などのデータ形式に変換する OCR ソフトウェアである．InftyReader は，科学技術情報のコンピュータ処理に関する研究・開発を目的とする非営利の任意団体である InftyProject[2] によって開発されたシステムであり，科学技術情報を容易に扱うためのユーザーインターフェースの実現や，視覚障がい者のための科学技術情報のアクセシビリティ向上を目的としている．
2. PhotoMath[3]
   PhotoMath は，算数や数学の学習支援を目的としたスマートフォンアプリケーションであり，数式にスマートフォンカメラを向けるだけで，その数式の解答や計算過程を導出する. 文書中の数式をコンピュータで計算・編集・検証できるような形式に変換する数式 OCR の技術が活用されている．
3. 数式検索 [4]
   数式検索は，検索したい数式と類似する構造または数学的な意味を持つ数式を，数式 OCR によって電子化し蓄積された論文やテキスト中から検索するようなシステ

ムである．数式検索のシステムを実現するための研究が，数式認識に関する研究とともに進められている．

このように，数式 OCR の技術は数学文書のアクセシブルな電子化や，数学の学習支援の用途のために実用化されつつある．

1.1.2　数式認識の課題

一般的に，文書中の数式を認識するためには，文書中から数式領域の検出，検出された数式領域中の文字・記号の切り出しと認識，数式構造の解析の処理が必要である [4]．文字・記号を正しく認識するためには，基本的にそれぞれの文字・記号を個別に切り出す必要がある．数式領域中の文字・記号の認識は，類似する形状の大文字と小文字 (C と c，S と s 等) の区別に加えて，数学的な意味が異なる立体と斜体 (C と C，s と s 等) も区別して行う必要があるが，個別に切り出された数式領域中の文字・記号に対しては約 97.7%の性能で正しく認識できることが報告されている [5].

正確に文字・記号を切り出せる文書が対象であれば，数式認識は実用可能な性能に達しているが，より高精度な数式認識のための課題として，接触記号の問題が内田らによって指摘されている [6].

1.2　本研究の目的

接触記号とは，印字する機器の性能や用紙の品質による滲み，または文書を画像化する際の解像度や適切でないしきい値による二値化などの要因により，隣接する文字・記号の一部分が，接触して構成される連結成分である (図 1.1).

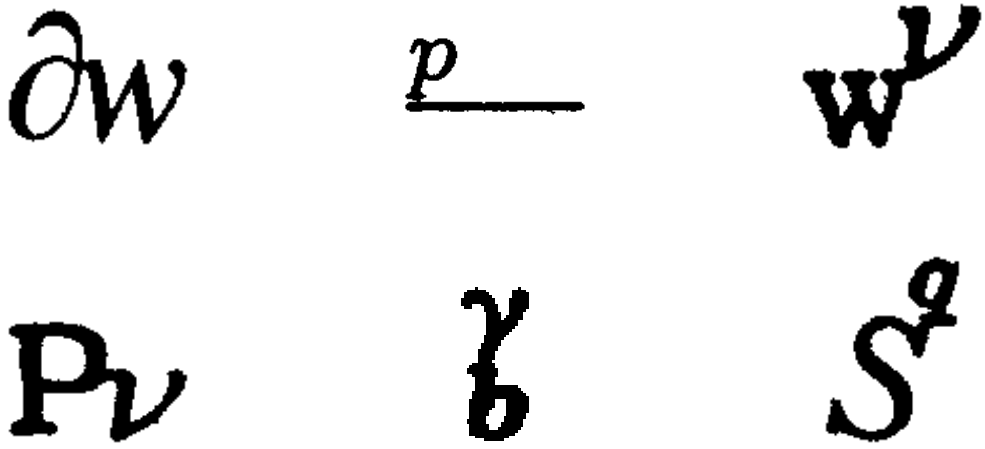
内田らは数学文書中のテキストと数式を解析し，テキスト中の接触記号と数式中の接触記号の特性を比較した [6]．それによると，テキスト領域の接触記号は基本的に水平方向にのみ接触し，外接矩形や垂直線により容易にセグメンテーションができる．また単語辞書を用いた校正によって接触する文字も正しく認識されることが多い．しかし，数式領域では定まった単語辞書が存在せず，数式の 2 次元構造により水平方向以外 (垂直方向や斜め方向) の接触も起こりうる (図 1.2)．そのため，テキスト領域の接触記号と比較すると，

図 1.1: 接触記号の例

数式領域の接触記号は分離・認識が困難な課題である．また，内田らは数式領域の接触記号を正しく分離・認識できなければ，後の数式構造の解析にも大きな悪影響を及ぼすとも指摘している．従って，数式領域における接触記号の分離手法の開発は，高精度な数式 OCR の実現において重要な課題と位置づけている．

公開されている数学文書用の OCR ソフトウェアである InftyReader[1] の開発グループにより接触記号分離に関する研究も行われていたが [7]，現在の InftyReader では鮮明に印刷された文書を対象としており，接触記号の分離・認識のための処理が実装されておらず，接触記号がある場合にはそれらを誤認識する。そのため認識前に手作業による文書画像の補正処理や，認識後に誤認識した文字・記号や数式構造の校正が必要となることが記されており，数式 OCR の接触記号への対応は試みられてはいるが，未だ実用に十分な手法が開発されておらず，困難な課題であると考えられる．

本研究では，より高精度な数式 OCR の実現のために，数式領域における接触記号の分離手法を提案する．本研究は，数式領域中から連結成分として切り出された文字・記号のうち，内田らの数学文書の解析結果に基いて数式領域中で発生率が高いとされた，2 つの文字・記号が接触している場合の接触記号を対象として，その接触記号を正しく文字認識できるように個別の文字・記号に分離する手法の開発を目的とする．



(a) 水平方向の接触　(b) 垂直方向の接触　(c) 斜め方向の接触

図 1.2: 数式の接触記号の例

第 2 章

関連する研究及び技術

2.1 接触記号分離に関する基本的方針

接触する文字・記号の分離に関する共通の方針として，接触記号分離のサーベイ論文 [8, 9] に “Dissection ”，“Recognition based segmentation ”，“Holistic ” の 3 種類の方針が挙げられている．

“Dissection ” は古典的な方針であり，幾何学的に共通する文字・記号らしさの特性に基いて接触記号を分離する手法である．例として，文字・記号に共通の横幅や高さのサイズや隣接する文字間のスペースの特徴によって分離を行う方法がある．

“Recognition based segmentation ” は分離記号仮定の生成と分離記号の決定の 2 つのステップからなる手法である．分離記号仮定の生成では，スライディングウィンドウ等によって接触記号の分離位置をずらしながら分離を行い，分離記号仮定を生成する．生成した分離記号仮定に対して文字・記号認識によって検証を行い，認識の結果から最良な分離の位置を選択することで接触記号を分離する手法である．認識に基づく分離方針の特徴は，文字・記号の認識のための学習サンプルと，分離記号仮定の生成と認識のための計算時間が必要となるが，性能の良い文字・記号の認識システムを用いれば高精度な分離が行える．

“Holistic ” は単語の認識に基いて分離を決定する手法である．あらかじめ学習された単語の情報に基いて文字列から文字を切り分け，接触する文字・記号も個別に分離を行うため，定まった単語や語彙を学習できる場合には有効な手法である．しかし，数式領域においては定まった単語辞書が存在しないため数式中の接触記号分離に対して “Holistic”な手法は適切でないと考えられる．

2.2 画像マッチングによる分離手法 [7]

野村らは，文書中の非接触記号をテンプレートとした画像マッチングによって同文書中の接触記号を分離する手法を提案した．野村らの手法は，文字認識の結果に基づいて文書中の連結成分すべてを接触記号のクラス X と非接触記号のクラス $\overline{X}$ に分類し，$\overline{X}$ に分類された連結成分をテンプレートとした画像マッチングによって X に分類された接触記号の分離を行う．数式の多様な接触（水平，垂直，対角方向）に対応するために，画像マッチングは接触記号の候補 $x \in X$ の左上，右上，左下，右下の四隅を調べる．画像マッチングにより一致する非接触記号 $y \in \overline{X}$ を x を構成する記号の一方と仮定し，さらに x から y と一致した部分を取り除いた差分画像と一致する非接触記号 $z \in \overline{X}$ を x を構成するもう一方の記号と仮定する．最後に，y と z を合成して接触記号 w を生成し，w を元の接触記号候補 x とマッチングさせて，x を構成する記号が y と z であることを検証する．

野村らの手法の特徴は，同じ文書中に出現する非接触記号をマッチングのテンプレートとして用いることで文書中の記号のフォントやサイズに合わせた分離を行っていることである．しかし，実際には接触記号を構成する文字や記号が同じ文書中に非接触記号として出現しない場合が多く，分離性能は約 51%であった．

2.3　四方向の直線による切断手法 [10]

Garain らは多因子解析によって接触記号の切断位置の候補を求め，求めた切断位置の候補に対して水平線，垂直線，± 45 度の対角線の四方向の直線によって接触記号を切断する手法を提案した．この手法は良い認識結果を得られるまで接触記号を分離する位置の選択を試みる．選択された分離位置を通る水平線，垂直線，対角線 (45 度 または -45 度) によって接触記号の連結成分を分離する．分離された記号候補は文字認識によって検証され，認識不可能 (棄却) であれば別の分離位置で直線による分離を繰り返す．

Garain らが提案する手法は独自に収集した 2853 個の接触記号のデータセットに対して約 96%と高い精度の接触記号分離性能を達成しており，また，3 つ以上の記号が接触する連結成分の分離にも対応できる．しかし，決められた四方向の直線のみによる分離を行っているため，図 2.1 の (a) に示すような接触記号に対しては，図 2.1 の (b) のような垂直線や+45 度の対角線による分離線では互いの記号を適切に分けられず，図 2.1 の (c) のような角度の分離線が必要となる．同様に，図 2.2 のような接触記号は四方向の直線で分離できないと思われるが，Garain らの論文ではこのような接触記号について言及されていない．

(a) 原画像　(b) 垂直線と+45 度の対角線　(c) 適切な角度の分離線

図 2.1: 四方向の直線で適切に分離できない例

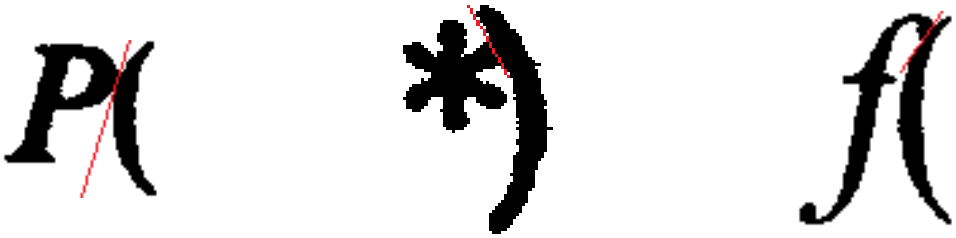


図 2.2: 四方向の直線で適切に分離できない接触記号

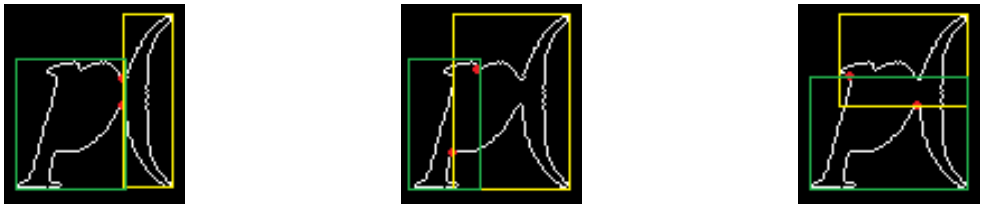
2.4　コーナー特徴点を結ぶ線分による分離手法 [11]

Zhang らは，接触記号の外側の輪郭線を抽出し，抽出した輪郭線上に検出される凹状のコーナー特徴点を用いて接触記号の分離線を求めた．この手法は，接触記号の接触箇所はストローク幅が小さくなり窪んだ状態になるという仮定に基づいた手法である．入力された接触記号 (図 2.3(a)) の連結成分の外側の輪郭線を抽出し (図 2.3(b))，Chetverikov らのコーナー検出手法 [12] によって凹状の特徴点を複数検出する (図 2.3(c))．検出された特徴点のうち任意の 2 点を通る直線を分離線の候補とする．求めた分離線の候補によって分離された記号に文字認識を行い適切な分離であるか検証を行うが，計算時間削減のために分離された 2 つの記号候補の重なり率によって文字認識を行う記号候補を削減している．図 2.4 の (a) に表す特徴点によって連結成分を分離すると，分離記号候補のそれぞれの外接矩形の重なり率は小さくなるが，図 2.4 の (b) や (c) に表す特徴点によって分離した場合には外接矩形の重なり率が大きくなる．重なり率が小さい順に 3 つの分離線を用いて分離した記号候補に対してのみ文字認識を行う．

Zhang らの手法は数式の接触記号を分離するための分離線の適切な角度を凹状のコーナー特徴点を用いて求めたものである．接触箇所の輪郭上に凹状のコーナー特徴点が検出できれば適切に分離が行えるが，図 2.5 に示すような接触記号は接触箇所が凹状になっておらず，図 2.5(a2) や (b2) のように接触箇所の輪郭上に特徴点が検出されないため適切な分離線を引くことができない．

(a) 原画像　(b) 接触記号の外側の輪郭線　(c) コーナー特徴点検出

図 2.3: Zhang らの手法



(a)　(b)　(c)

図 2.4: 重なり率による分離記号候補の削減

(a1) 原画像　(a2) コーナー特徴点の検出例

(b1) 原画像　(b2) コーナー特徴点の検出例

図 2.5: 接触箇所に特徴点が検出されない接触記号の例

第 3 章

提案手法

3.1 提案手法の概要

本研究では，数式領域において発生する接触記号に対して高い精度の分離を行うために，2 章の「関連する研究及び技術」で述べた認識に基づく分離 (Recognition based segmentation) を採用し，数式領域中の接触記号から分離記号候補を生成して，生成した分離記号候補の中から文字認識によって最良な分離記号を決定することで接触記号の分離と認識を行う.

従来の研究では，数式領域領域における接触記号の特徴である水平方向以外 (垂直方向や斜め方向) の接触に対応するために，野村ら [7] による接触記号の四隅に対する画像マッチングで重なる領域を除去する分離記号候補の生成や，Garain[10] らによるあらかじめ決めた四方向の直線による分離，また Zhang[11] らによるコーナー特徴点に基いて求めた直線による分離によって分離記号候補を生成する手法が提案されていた.

本研究で提案する手法は，従来手法のように太さのある接触記号の連結成分を直接分離せず，接触記号の連結成分から幅が 1 画素の芯線を抽出し，その芯線に対して分割を行った後，分割された各芯線を元の連結成分の距離変換画像と重ね，芯線上の距離値を用いて連結成分を復元することで，近似的に太さをもった接触記号に分離する (方法の詳細は 3.4 で述べる). 接触記号の連結成分から抽出した芯線の分割は，直線による分離を必要とせず，芯線上の画素 (点) の除去によって連結成分を分けられる. 芯線分割による接触記号の分離は，接触記号の連結成分を直線で分離する際に考慮すべき適切な分離線の角度を求める必要がなく，また適切でない角度の分離線に起因する接触記号の誤った分離を防ぐことができる.

文字認識による最良な分離記号の決定のために，特徴ベクトルとして分離記号候補から抽出される濃度こう配特徴 [13] を，識別関数として Modified Quadratic Discriminant Function(MQDF)[14] を採用した文字認識システムを用いる. 本システムによって分離さ

れた分離記号候補の全てを対象に文字認識を行い，認識結果の字種と，その相違度 (値が小さいほどその文字らしい) を算出する．さらに相違度の和 (評価値) が最小となる分離記号の組み合わせを分離記号として決定する．

3.2　提案手法の処理手順

提案手法の処理手順を次に示す．

1. 接触記号の入力
2. 前処理
3. 分離記号候補の生成
   （a）ストローク芯線の抽出
   （b）芯線の分割
   （c）分割された芯線からの連結成分復元
4. 文字認識の評価値に基づいた分離の決定
5. 分離記号の出力

本手法による入出力の例を図 3.1 に示す．各処理の詳細について以降に述べる．

(a) 入力　(b) 出力 1　(c) 出力 2

図 3.1: 入出力の例

3.3 前処理

接触記号は，本来離れて印字されるはずの文字・記号の一部分が，印字する機器の性能や用紙の品質，また画像化する際の解像度や適切でないしきい値による二値化などの要因によって接触したものである．接触が軽度であれば，2 つの文字・記号の接触箇所には小さな孔が存在する場合がある．

図 3.2 の (a) ように接触箇所に小さな孔が存在する接触記号から芯線を抽出すると，(b) のような芯線が得られ，接触箇所が複数の線によって連結され，後の芯線分割に悪影響を及ぼす．

前処理では，入力された接触記号の連結成分にしきい値以下の面積の孔があれば，その孔を埋める処理を行う．図 3.3 の (a) に対して前処理を行い (b) のように孔を埋めることで，芯線を抽出した際に (c) のような接触箇所が一本の線で連結されるような芯線を得られる．

(a) 原画像　　(b) 芯線の抽出例

図 3.2: 接触箇所に孔が存在する接触記号

(a) 原画像　　(b) 前処理後の画像　　(c) 芯線の抽出例

図 3.3: 前処理の例

3.4　分離候補の生成

芯線分割と連結成分の復元による分離候補生成の手順について述べる．

3.4.1　ストローク芯線の抽出

前処理を行った連結成分に対して Hilditch の細線化アルゴリズムを適用し，ストローク芯線を 4 連結で抽出する．Hilditch の細線化アルゴリズムによる芯線の抽出例を，図 3.4 に示す．また前処理後の連結成分に 8 近傍の距離変換を行い，芯線から連結成分復元のために用いる距離画像を生成する．

(a) 原画像　　(b) 抽出された芯線

図 3.4: 芯線抽出の例

3.4.2　芯線の分割

抽出された芯線の全画素を探索し，注目画素において 1 つの連結成分である芯線を 2 つの連結成分に分割できるか判定を行う. 芯線を校正する画素の値を 1，背景を構成する画素の値を 0 として，注目画素の画素値を 1 から 0 にした後にラベリングを行い連結成分の個数を調べる. 連結成分の個数が 2 つであれば，注目画素において 1 つの連結成分である芯線を 2 つの連結成分に分割可能と判断できる.

図 3.5 に芯線の分割の例を示す．図中の値は各画素の値を表す. 芯線の連結成分を構成する画素の値を 1，背景を構成する画素の値を 0 としている.

0	0	0	0	0	0	0
0	1	1	1	1	1	0
0	0	0	0	0	0	0

(a) 芯線分割前

0	0	0	0	0	0	0
0	1	1	0	1	1	0
0	0	0	0	0	0	0

(b) 芯線分割後

図 3.5: 芯線の分割例

図 3.5 の (a) において黄色で示す画素を注目画素とした時, 図 3.5 の (b) のように注目画素の値を 1 から 0 にする. この操作によって 1 つの芯線の連結成分を注目画素で 2 つの連結成分に分割する.

元の芯線を構成する画素のうち，1 つの芯線を 2 つの連結成分に分離可能である画素の全てに対して 2 つの連結成分への分割を行う.

2 画素による芯線分割

図 3.6 に示すように，接触記号の接触箇所が 2 箇所ある場合には，1 画素による分割では芯線を適切な位置で分割できない. このような場合に対応するために，注目画素の除去によって芯線を分割できなければ，注目画素と同じ X 座標，または Y 座標に位置する芯線上の画素も除去することで芯線分割を行う.

図 3.7 に 2 画素による分割の例を示す．図 3.7 の①で示した画素を注目画素として除去した場合，X 座標が同じである②で示す画素も同様に除去し，芯線を 2 分割する.

(a) 原画像　　(b) 抽出された芯線

図 3.6: 芯線を 1 画素で分割できない接触記号の例

(a)　　(b)

図 3.7: X 座標が同じである 2 画素による分割例

3.4.3 芯線からの連結成分の復元

分割された芯線から太さのある連結成分を復元する．連結成分の復元には分割された芯線画像と，前処理直後の画像に対して 8 近傍の距離変換を行って得られる距離画像を用いる．前処理直後の画像に 8 近傍の距離変換を行い，画像中の接触記号を構成する連結成分の各画素から背景画素までの最短距離値がその画素における距離値として求められる．分割された芯線の画像と距離画像の座標を照らし合わせて，芯線を構成する各画素に対応する距離値を得る．

1. 分割された芯線の画像にラスタ走査を行い，芯線を構成する画素 (画素値 1) が見つかれば注目画素 $p(x, y)$ とする．
2. 注目画素の座標 (x, y) に対応する距離値 n を得る．
3. 注目画素 $p(x, y)$ を中心に以下の式によって太さのある連結成分を得る．

$$\sum_{j=y-n}^{y+n} \sum_{i=x-n}^{x+n} p(i, j) = 1 \tag{3.1}$$

4. 芯線画像の全てを走査するまで手順 1∼3 を繰り返す．
5. 手順 1∼4 において処理された画像に対して，前処理直後の画像で背景 (画素値 0) である座標の画素値を 0 に更新し，芯線を抽出する前の連結成分よりも余分に復元した部分を背景に戻す．

以上によって分離記号候補を生成する．

芯線分割と連結成分の復元による分離記号候補生成の流れを図 3.8 に示す．図 3.8 の赤い矢印で示す注目画素において芯線を 2 つの連結成分に分割した時に図 3.8 右のような分離記号候補が生成される．

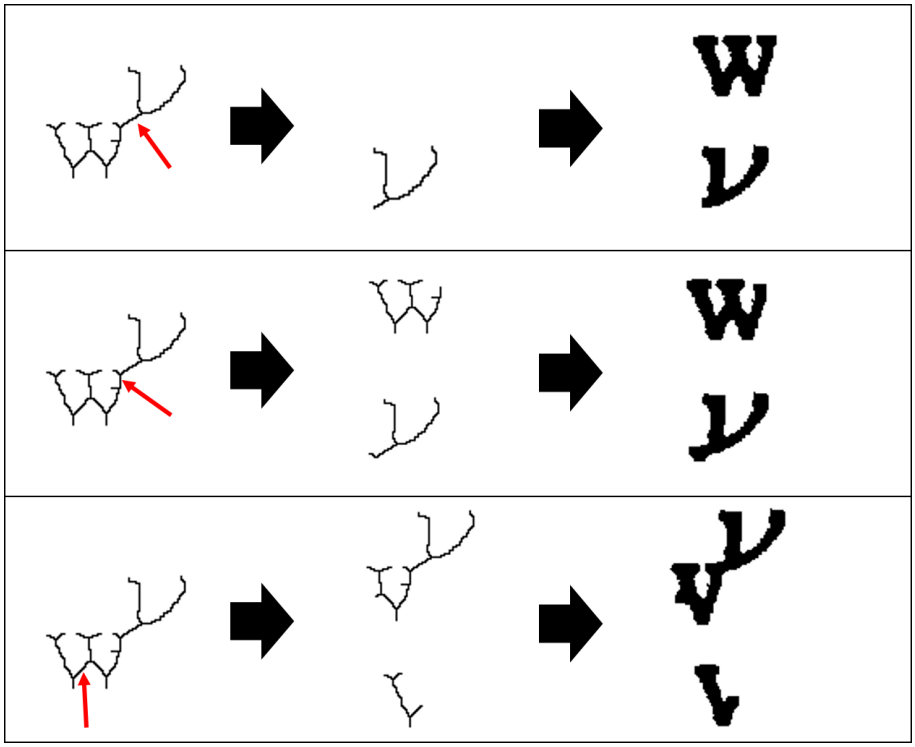


図 3.8: 芯線分割と連結成分の復元による分離記号候補の生成例

3.5　文字認識の評価値に基づいた分離の決定

分離記号候補として生成された連結成分のそれぞれに対して文字認識を行い，分離記号候補と認識結果のクラスとの相違度の和を評価値として最終的な分離結果を決定する．

3.5.1　使用する文字認識システムの概要

特徴ベクトル

特徴ベクトルは，濃度こう配特徴 [13] を用いて抽出する．1 画像から，400 次元の濃度こう配特徴を抽出する手順を示す．

1. 入力画像のサイズを 65×65(pixel) に正規化する．
2. 2×2 の平均値フィルタを 5 回適用して平滑化された濃淡画像を得る．
3. 濃淡画像の平均が 0，分散が 1 となるように画像を正準化する．
4. Roberts フィルタを適用し，こう配の強度と向きを求め，向きを $\frac{\pi}{16}$ 刻みで 32 向きに量子化する．
5. 画像中の文字・記号の外接枠を 81(9×9) の小領域に分割し，各領域内で量子化した 32 の向きごとにこう配の強度を加算することで局所方向ヒストグラムを作成する．
6. 加重フィルタにより，32 向きのヒストグラムを 16 向きに圧縮する．さらに方向別に 2 次元ガウスフィルタを施して領域数を 9×9 から 5×5 に圧縮して，400 次元のこう配方向ヒストグラムを得る．
7. 抽出した特徴ベクトルの各成分に対して変数変換 [15] を行い，特徴量の分布を正規分布に近づける．

抽出した 400 次元の濃度こう配特徴ベクトルに対して主成分分析による特徴選択を行い次元数を 250 に削減する．以上の処理によって得られた 250 次元の特徴ベクトルを文字認識に用いる．

識別関数

識別関数には Modified Quadratic Discriminant Function(MQDF)[14] を用いる．MQDF は分布パラメータのうち母集団の共分散行列が未知の正規分布に対する最適識別関数から導出された近似式で，識別精度を損なうことなく識別時の計算量と記憶容量を $O(n^2)$ から $O(kn)$ に削減できる特徴がある．学習サンプル数に対して，特徴ベクトルの次元数を過度に増やすと性能が低下するピーキング現象を抑える効果があり，共分散行列の推定誤差に起因する性能低下も少ない．本研究では，各パラメータは数式中の単文字・記号に対しての認識実験によって最適であった $\alpha = 0.9$，$k = 38$ を用いる．

$$g(X) = \frac{1}{\alpha\sigma^2}\left[\|X-M\|^2 - \sum_{i=1}^{k}\frac{(1-\alpha)\lambda_i}{(1-\alpha)\lambda_i+\alpha\sigma^2}\left\{\Phi_i^{\mathrm{T}}(X-M)\right\}^2\right] + \sum_{i=1}^{k}\ln\{(1-\alpha)\lambda_i+\alpha\sigma^2\} \tag{3.2}$$

X : 入力文字の特徴ベクトル
M : 母集団の平均ベクトル
λ_i : 標本共分散行列の第 i 固有値
Φ_i : 標本共分散行列の第 i 固有ベクトル
k : 識別に用いる固有ベクトル数
σ^2 : 特徴ベクトル X の事前確率分布を球状分布と仮定した場合の分散
α : σ^2の信頼度を表す定数

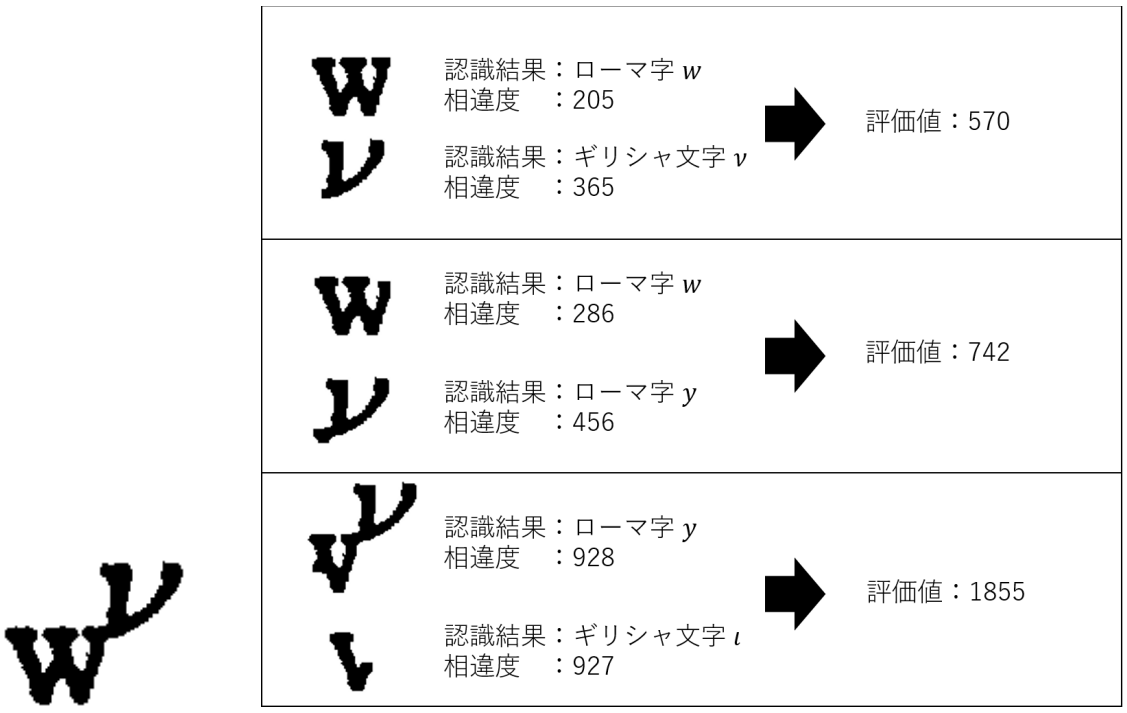
MQDF では，$g(X)$ が最小となる文字・記号クラスが認識結果となる．分離記号候補のそれぞれの MQDF による認識結果の $g(X)$ の和を分離記号の決定のための評価値として用いる．

3.5.2　分離の決定

前述した文字認識システムを用いて，生成した分離記号候補の中から最良な分離記号を 1 組選択し，分離記号の決定を行う．本手法によって算出される評価値は学習されている文字・記号種との相違度を表すため，値が小さいほど認識したクラスの文字・記号らしいと言える．

本手法では，2 つに分離された分離記号候補のそれぞれから算出された相違度の和を分離候補ごとに求め，分離候補中で相違度の和である評価値が最小となる分離記号の組を，最良の分離結果として出力する．

評価値に基づいた分離記号の決定例を図 3.9 に示す．図 3.9(b) では上の w と ν と認識された場合に評価値が最小となり，最終的な分離結果として決定される．



(a) 入力画像　　(b) 分離記号候補と評価値の算出例

図 3.9: 評価値に基づく分離記号の決定例

図 3.10 は接触記号の芯線の分割位置と評価値の対応を色で表したものである．図の赤で表される位置で芯線を分割した場合には評価値が小さく，色が青くなるほどその位置で分割した場合に評価値が大きくなることを表している．

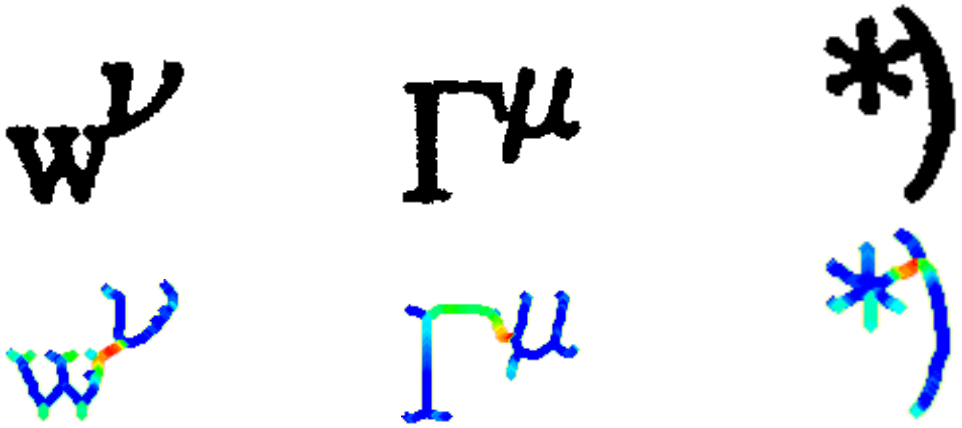


図 3.10: 芯線の分割位置と評価値の対応

第 4 章

性能評価実験

4.1 使用するデータベース

性能評価実験のために，公開データベースである “InftyCDB-1 ”, “InftyCDB-3-A ” [16] を使用する．InftyCDB は科学文書用の OCR ソフトウェアの研究・開発のために公開されているデータベースである．InftyCDB-1 には，30 の英文数学論文，全 476 ページから採取された 688,580 個の文字・記号の画像と正解文字コードが収納されており，収録されている単語数は 108,914，数式数は 21,056 である．図 4.1 に InftyCDB-1 データベースに収録されている画像の例を示す．

InftyCDB-3-A は英数字と数学記号の単文字画像データベースである．異なる出版社の本や論文，和書の数式，出版社の印字見本，PC に含まれる内蔵フォント，LaTeX のフォントなど，多様なソースから採取された 188,752 の単文字・記号の画像が含まれている．収録されている字種は数字，ローマ字，ギリシャ文字，数式記号 (図 4.2) など計 384 クラスであり，数字とローマ字は立体と斜体で異なるクラスとして用意されている (図 4.3).

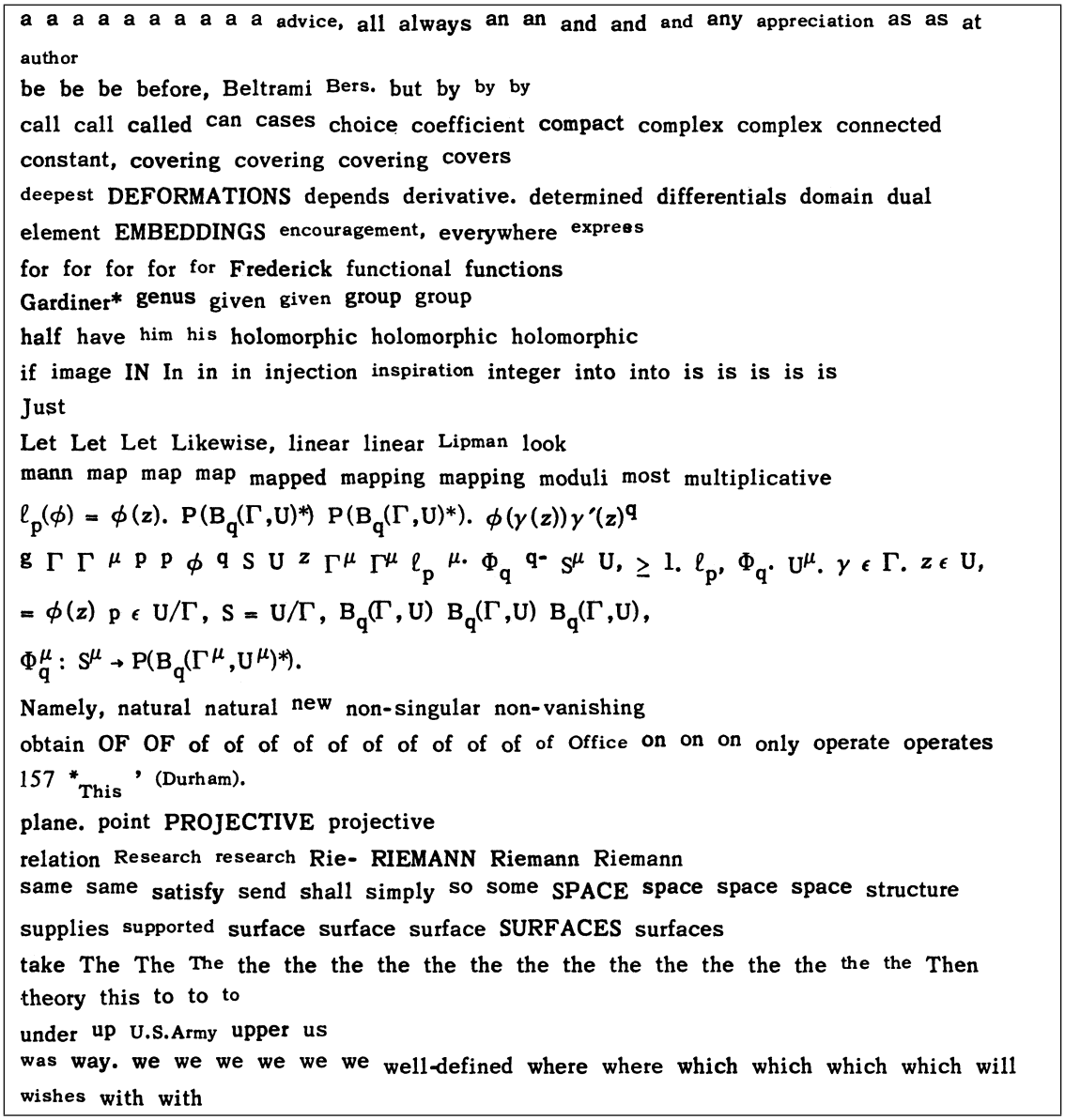


図 4.1: InftyCDB-1 データベース

図 4.2: InftyCDB-3-A データベースに含まれる文字・記号の例

(a1) 立体大文字 C　(a2) 斜体大文字 C　(b1) 立体小文字 s　(b1) 斜体小文字 s

図 4.3: InftyCDB-3-A に含まれる立体と斜体の例

4.1.1 評価用データセット

InftyCDB-1 には接触記号が含まれており，アノテーションとして与えられている情報から接触記号を抽出できる．アノテーションを用いて抽出した，InftyCDB-1 データベースの数式中の 2 記号からなる接触記号である連結成分の 627 個を評価実験に用いる．評価用データの例を図 4.4 に示す．

図 4.4: 評価用データセットの接触記号例

4.1.2 学習用データセット

学習用データセットには，単文字・記号のデータベースである InftyCDB-3-A 中の画像を使用する．本実験では InftyCDB-3-A に含まれる 384 クラスの字種をそのまま学習に用いる．1 クラスあたりの学習用データ数は最大 500 画像とする．

4.2 評価方法

入力した 2 記号からなる接触記号を提案手法によって分離し，2 記号のそれぞれを正しい記号クラスに識別すれば，その接触記号に対して分離成功とする．分離成功率の計算式を以下に示す．

$$\text{分離成功率} = \frac{\text{分離に成功した接触記号の数}}{\text{入力する接触記号の総数}}$$

4.3　数式中の接触記号の分離実験

提案手法によって評価用データセット中の接触記号の分離実験を行い，提案手法の分離性能を評価する．

4.3.1　結果と考察

実験結果は，627 個中 594 個の連結成分に対して分離成功し，分離成功率は 94.7%であった．分離成功例を図 4.5 に示す．提案手法の芯線分割によって，水平以外の方向に接触する数式の接触記号についても分離して正しく認識することができた．

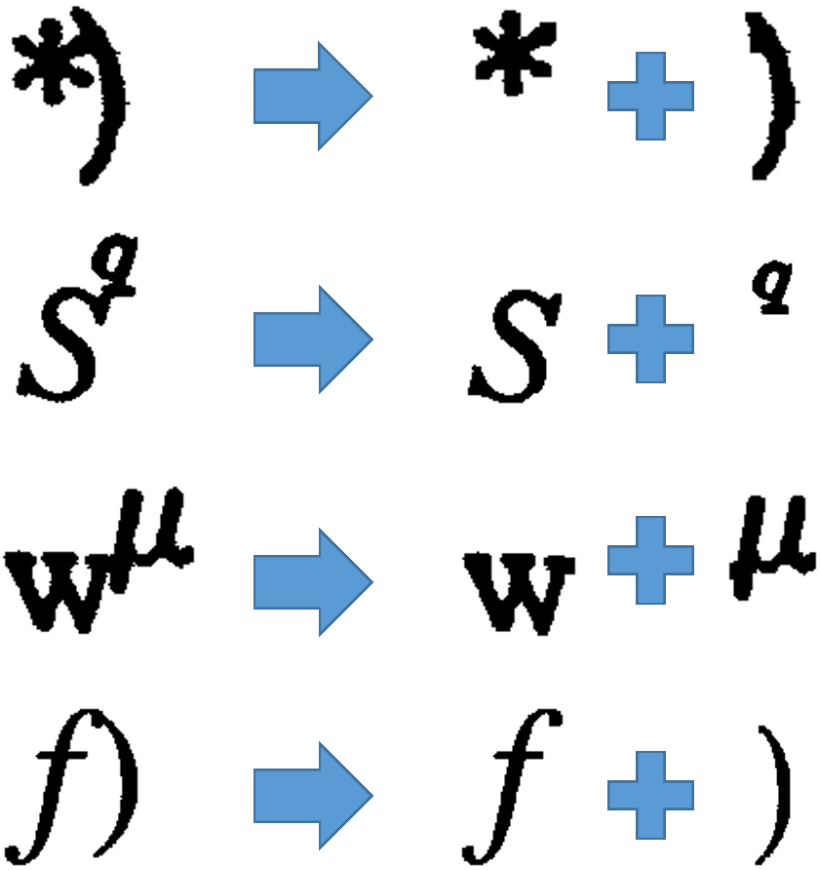


図 4.5: 分離成功例

分離失敗の原因について以降に述べる．

芯線分割の失敗

提案手法は，芯線分割と太さのある連結成分の復元によって接触記号を分離しているが，適切に芯線を分割できず，望ましい分離記号候補を生成できず失敗する例があった．芯線分割を失敗する例を図 4.6 に示す．

(a)2 箇所で接触する接触記号　　(b) 接触が重度の接触記号

図 4.6: 芯線分割を失敗する接触記号の例

図 4.6 の (a) は隣接する文字・記号が 2 箇所で接触する例である．このような例は，3.4.2 で述べた，X 座標または Y 座標が同じである 2 画素を用いた芯線分割によって分割できる場合もあるが，図 4.6(a) のように接触箇所が 2 箇所あり，かつ 2 つの接触箇所が斜めに位置する場合には正しい分離記号候補を生成できない．2 箇所が接触する接触記号の例は 4 例あり，1 画素のみで分割を行う場合には 4 例全てを分離失敗するが，2 画素による分割でも失敗する例は図 4.6(a) に示した 1 例のみであった．

図 4.6 の (b) は接触が重度であるため，芯線が分割に適さない形状で抽出される例である．このような失敗例は図 4.6(b) の 1 例のみであった．

図 4.7 に示すように，接触が重度である (a1) の接触記号から (a2) のような芯線が抽出され，2 つの文字・記号の接触部分の芯線が重なってしまい，適切に芯線を分けることができず失敗となった．接触が軽度である図 4.7(b1) の接触記号からは (b2) のような芯線が抽出され，正しく分離・認識が行えた．

(a1) 接触が重度の接触記号

(a2) 抽出された芯線

(b1) 接触が軽度の接触記号

(b2) 抽出された芯線

図 4.7: 接触の度合いと抽出される芯線の比較

ノイズ等による失敗

ノイズや汚れ，滲みによって接触記号の形状が崩れるため分離を失敗する例が 3 例あった．図 4.8 に例を示す．これらの接触記号は芯線分割によって分離記号候補を生成しても文字認識を正しく行えなえず，正しく分離・認識をするためにノイズ除去のような特別な処理が必要であると考える．

図 4.8: ノイズ等による失敗例

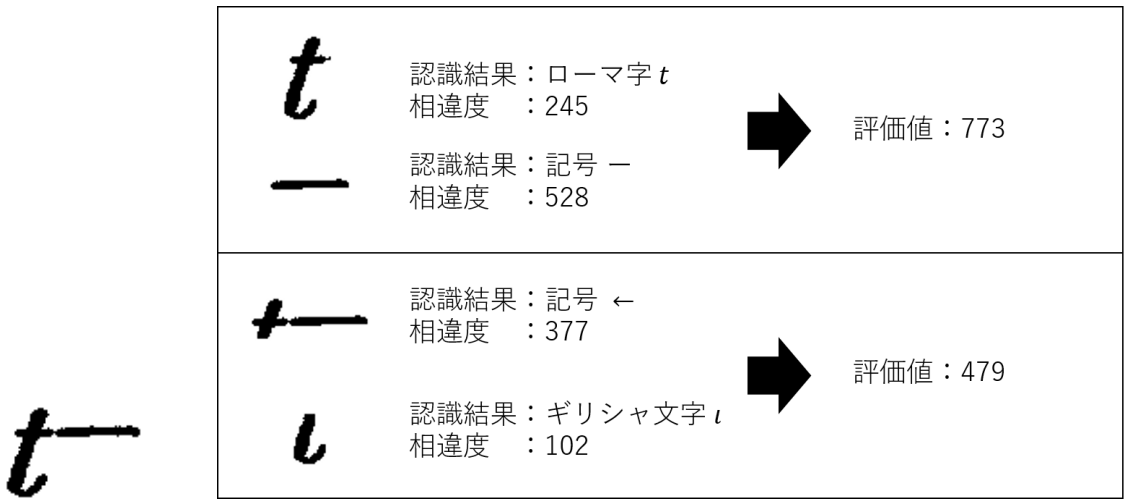
文字認識による分離記号の決定の失敗

分離記号の決定の失敗は，入力された接触記号に対して適切な位置で分離記号候補が生成され，それぞれの文字・記号を正しく認識されたが，誤った位置で分離した記号候補の形状が学習されているいずれかの字種の形状と類似しているため，誤った分離記号候補が選ばれたことで失敗となった．InftyCDB-1 から抽出した数式の接触記号に関しては分離記号の決定の失敗が最も多く，28 例の接触記号で失敗した．

図 4.9 に分離記号の決定の失敗例を示す．(a) の接触記号に対して (b) のような認識結

果と評価値が得られ，(b) の上の例のように正しく分離・認識された分離記号の評価値より，誤った分離・認識を行った下の例の方が評価値が良くなり失敗となった．

このような例に対応するためには文字認識システムの改善が必要であると考えられる．文字認識システムに学習されている文字・記号のフォントと認識対象の文字・記号のフォントが違えば，同じクラスの文字・記号種に対しても相違度が高くなり，分離記号候補の決定のための評価値も悪くなる．文字認識システムの精度向上や，フォント推定とそれへの適応性をもつ文字認識システム [17] を用いることで，分離記号決定の失敗を改善可能であると考えられる．



(a) 入力画像　　(b) 分離記号の決定例

図 4.9: 分離記号の決定の失敗例

4.3.2 実行時間

提案手法の実行時間について述べる．実験の実行環境は表 4.1 に示す．

今回の実験では，芯線分割と連結成分の復元により 627 個の接触記号から 96,088 個の分離記号候補が生成され，分離記号候補の生成に要した実行時間は 50.4[sec] であった．また，生成した 96,088 個の分離記号候補に対して文字認識を実行し，入力された接触記号に対して最良な評価値の分離記号を決定するのに要する実行時間は 1,241.6[sec] であった．実行時間の合計は 1,292[sec] であり，1 画像あたり 2.07[sec] の時間が必要であった．

提案手法の実行時間は，生成される分離記号候補の数によって増減する．入力された接触記号のサイズや形状に依存して芯線が抽出され，芯線を構成する画素数が多いほど，生成される分離記号候補も多くなる．

表 4.1: 実行環境

CPU	Intel Core i7-2600 3.4GHz
OS	CentOS 6.8
メモリ	4 GB

第 5 章

結論

5.1 本研究のまとめ

本論文では，数式 OCR において問題となる接触記号のうち，隣接する 2 記号が接触する場合の分離手法を提案した．提案手法は接触記号である連結成分の芯線を分割することで，分離に適した分離線の方向を考慮せずに分離を行える接触記号の分離手法である．最適な分離の位置の決定には MQDF による文字認識の評価値を使用した．

本手法は，公開データベースである InftyCDB-1 中の，2 記号が接触する接触記号である連結成分 627 個に対して，94.7%の分離成功率を得られ，水平以外の方向で接触する数式の接触記号にも有効な手法であることを示した．

5.2 今後の課題

5.2.1 認識システムの高精度化

InftyCDB-1 の接触記号に対しての分離実験で，最も多い失敗原因となった文字認識による分離記号決定の誤りを改善するために，文字認識システムの高精度化が必要である．

文献 [17] の研究では，ウェブ上で公開されている多種類のフォントの文字に対して，その輪郭線の角点等の特徴量の (x, y) 座標からつくられた特徴ベクトルを求め，特徴空間上で主成分分析を行った結果を用いて，人間の可読範囲を特徴空間上に設定し (心理的計測による)，その空間内の多様な点に対応する多数の文字を発生させて学習文字をつくる．従って非常に多種類のフォントに対応した認識システムとなることが期待されるとともに，フォントの適応機能をもたせることにより，極めて高精度な認識性能が得られることを実験により確認している．

5.2.2 3 つ以上の文字・記号の接触

本研究では，数式領域において発生率が高いとされる 2 つの文字・記号の接触による接触記号のみを分離対象としたが，3 つ以上の文字・記号が接触することも考えられる．より高性能な接触記号分離のためは 3 つ以上の文字・記号の接触への対応が課題である．

5.2.3 分離記号候補の削減

本手法では，芯線分割によって，1 つの接触記号から平均して 100∼200 個の分離記号候補を生成したが，最終的な分離記号として決定されるのは 1 組 (2 個) の分離記号のみである．分離記号候補の生成には実行時間は 1 画像につき平均 0.08[sec] 程であるが，分離記号候補の全てに文字認識を行うための計算時間が膨大となる．

この問題を解決するために生成された分離記号候補を削減し，文字認識を行う分離記号候補を減らす手法の検討が必要である．

付録 A

付録

A.1　研究用ディレクトリ・プログラム

研究に用いたプログラム，データセットはヒューマンインターフェース研究室サーバー内の以下のディレクトリ下に置く．

- “/net/xserve0/users/kawaguti/research/”

また，research ディレクトリ内に存在するディレクトリとファイルの詳細を以下のファイルに示す．

- “/net/xserve0/users/kawaguti/research/readme.txt”

研究用プログラムをビルドするために, 画像処理ライブラリの OpenCV(“https://opencv.jp/”) のインストールが必要である．本研究では，OpenCV2.4.9 のバージョンを用いてプログラムを作成している．

A.2　修論発表会プレゼンテーション

末尾に，平成 28 年度修論発表会において発表したプレゼンテーションのスライドを掲載する．

謝辞

本研究を進めるにあたり，様々な方のご協力をいただきました．研究に対する基本姿勢や様々なアイデア，アドバイス，専門知識と技術を御教授下さった三宅康二名誉教授，木村文隆名誉教授，若林哲史教授，大山航助教に深く感謝します．また，日頃の研究生活を支えていただいた田中みゆき事務主任，共に過ごしたヒューマンインターフェース研究室の学生の皆様，6 年間に渡る大学，大学院の生活をサポートしていただいた家族に対し感謝を表し，本論文の結びといたします．

参考文献

[1] “InftyReader” http://www.sciaccess.net/en/InftyReader/ (2017 年 2 月 1 日)

[2] “Infty Project” http://www.inftyproject.org/en/index.html (2017 年 2 月 1 日)

[3] “photomath” https://photomath.net/en/ (2017 年 2 月 1 日)

[4] Richard Zanibbi and Dorothea Blostein: “Recognition and retrieval of mathematical expressions”, International Journal on Document Analysis and Recognition (IJDAR) 15.4 (2012): 331-357.

[5] Christopher Malon, Seiichi Uchida and Masakazu Suzuki: “Mathematical symbol recognition with support vector machines”, Pattern Recognition Letters 29.9 (2008): 1326-1332.

[6] Seiichi Uchida, Akihiro Nomura and Masakazu Suzuki: “Quantitative analysis of mathematical documents”, International Journal of Document Analysis and Recognition (IJDAR) 7.4 (2005): 211-218.

[7] Akihiro Nomura, Kazuyuki Michishita, Seiichi Uchida and Masakazu Suzuki: “Detection and segmentation of touching characters in mathematical expressions”, Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on. IEEE, 2003.

[8] Casey Richard G and Eric Lecolinet: “A survey of methods and strategies in character segmentation”, IEEE transactions on pattern analysis and machine intelligence 18.7 (1996): 690-706.

[9] Saba Tanzila, Ghazali Sulong and Amjad Rehman: “A survey on methods and strategies on touched characters segmentation”, International Journal of Research and Reviews in Computer Science 1.2 (2010): 103-114.

[10] Garain Utpal and B. B. Chaudhuri: “Segmentation of touching symbols for OCR of printed mathematical expressions: an approach based on multifactorial analysis”, Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on. IEEE, 2005.

[11] Zhang Dong-Yu, Xue-Dong Tian and Xin-Fu Li: “An Improved method for segmentation of touching symbols in printed mathematical expressions”, Advanced Computer Control (ICACC), 2010 2nd International Conference on. Vol. 2. IEEE, 2010.

[12] Chetverikov Dmitry: “A simple and efficient algorithm for detection of high curvature points in planar curves”, International Conference on Computer Analysis of Images and Patterns. Springer Berlin Heidelberg, 2003.

[13] 澤和宏, 若林哲史, 鶴岡信治, 木村文隆, 三宅康二: “こう配特徴ベクトルと変動吸収共分散行列による手書き漢字認識の高精度化”, 電子情報通信学会論文誌 D 84.11 (2001): 2387-2397.

[14] Fumitaka Kimura, Kenji Takashina, Shinji Tsuruoka and Yasuji Miyake: “Modified quadratic discriminant functions and the application to Chinese character recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1987): 149-153.

[15] Keinosuke Fukunaga: “Introduction to statistical pattern recognition”, Academic press, 2013.

[16] Masakazu Suzuki, Seiichi Uchida and Akihiro Nomura: “A ground-truthed mathematical character and symbol image database”, Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on. IEEE, 2005.
http://www.inftyproject.org/en/database.html (2017 年 2 月 1 日)

[17] 一柳仁志: “擬似フォントを利用した数式画像中の文字・記号の高精度認識”, 中部大学大学院工学研究科 平成 19 年度修士論文