卒業論文

日本語文中の英文字の認識手法の開発

東北大学工学部情報工学科４年
中川 修司

目 次

第 1 章 序論 **1**
1.1 本研究の背景 1
1.2 本研究の目的 2
1.3 本論文の構成 2

第 2 章 文字切り出しとその問題点 **3**
2.1 はじめに 3
2.2 基本的な文字切り出し法 3
2.2.1 射影 (投影) 分布を用いる方法 4
2.2.2 連結領域に着目する方法 4
2.2.3 その他 4
2.3 基本的な文字切り出し法の問題点 4
2.3.1 文字切り出しにおける日本語の特徴上の問題点 5
2.3.2 文字切り出しにおける英語の特徴上の問題点 5
2.3.3 対策 6
2.4 日本語と英語が混在している文書の切り出し法 7
2.5 まとめ 7

第 3 章 日本語文中からの英文字抽出 **8**
3.1 はじめに 8
3.2 文書画像入力 10
3.3 射影分布による文字要素の決定 10
3.4 平均文字幅、平均ピッチの決定 11
3.4.1 文字幅と文字ピッチ 11
3.4.2 一行あたりの文字幅、文字ピッチの分布 12
3.4.2.1 実験方法 12
3.4.2.2 調査対象 12
3.4.2.3 結果 12
3.4.3 平均文字幅、平均文字ピッチの推定 14
3.5 文字要素の評価 14
3.5.1 スコア 14

3.5.2 スコアのフィルタリング 17
3.6 英文字領域の抽出 18
3.6.1 英文字領域の判定 18
3.6.2 英文字らしさの閾値の決定 18
3.6.2.1 決定方法 18
3.6.2.2 結果 18
3.6.3 英文字抽出結果の検討 20
3.6.3.1 英文字の抽出率の検討 20
3.6.3.2 日本語の誤抽出の検討 21
3.6.3.3 抽出成功例の分析 21
3.7 まとめ 21

第 4 章 英文字抽出の有効性の検証 **23**
4.1 はじめに 23
4.2 英文字領域用の切り出し 23
4.2.1 切り出しの手順 24
4.2.2 分離処理 24
4.2.3 本切り出し法の評価 25
4.2.3.1 目的 25
4.2.3.2 比較方法 25
4.2.3.3 結果 25
4.3 英文字専用辞書の作成 26
4.3.1 英文字専用辞書の構成 26
4.3.2 英文字専用辞書の評価 28
4.3.2.1 目的 28
4.3.2.2 実験方法 28
4.3.2.3 結果 28
4.4 英文字抽出の有効性の評価 29
4.4.1 実験方法 29
4.4.2 実験結果 29
4.5 まとめ 30

第 5 章 結論 **31**
5.1 結論 31
5.2 今後の課題 31
5.2.1 英文字抽出の高精度化 31
5.2.2 文字切り出し法の強化 32
5.2.3 英文字用辞書の充実 32

謝辞 **34**

参考文献 **35**

図目次

2.1 射影 (投影) 分布を用いた文字切り出しの例 4
2.2 分離文字の例 5
2.3 文字の接触の例 5
2.4 英語の一般文書 (不定ピッチ) の例 5
2.5 英語の一般文書 (固定ピッチ) の例 5
2.6 kerning の例 6
2.7 イタリック体の例 6
2.8 ligature の例 6

3.1 英文字が混在する日本語文書例 8
3.2 英文字抽出の概略 9
3.3 文字要素決定までの過程 10
3.4 文字要素の外見情報の定義 11
3.5 日本語のみの行イメージの文字幅、文字ピッチの度数分布例 13
3.6 英語のみの行イメージの文字幅、文字ピッチの度数分布例 13
3.7 日本語と英語が混在している行イメージの文字幅、文字ピッチの度数分布例 13
3.8 文字要素自身のスコアリングの具体例 16
3.9 長さ 5 文字要素の方形フィルタ 17
3.10 スコアリングの様子 19
3.11 抽出できなかった英文字の例 20
3.12 日本語の誤抽出の例 21

4.1 辞書の学習サンプルに用いたフォント 27

5.1 ガウシアン的フィルタの 1 例 32

表目次

3.1 いろいろな閾値での英文字領域の抽出率 18

4.1 斜め切り出しの角度設定とその実際の角度 25
4.2 切り出し法の性能比較結果 26
4.3 英文字記号1セット一覧 26
4.4 本辞書に組み入れた連字の一覧 28
4.5 英文字専用辞書の性能比較結果 28
4.6 英文字抽出の有効性の検討のための認識実験の結果 29

第1章

序論

1.1 本研究の背景

計算機の発達、普及に伴い、人間の能力を機械(計算機)によって実現しようという努力が現在も行われている。中でも、人間の文字を読み取るという能力の実現である文字認識技術については古くから盛んに研究が行われている。その理由の1つとしては、文字が1文字に1つの概念が対応し、記録性もよく、我々にとって身近なものであるため、研究対象になりやすいことが挙げられる。また、もう1つの理由としは、文字を読み取る技術が計算機へのデータ入力の主流であるキーボードを介した作業を省力化、効率化するための装置の開発に結びつき、社会的なニーズも強かったことが挙げられる。

当初、光学文字読取装置(Optical Character Reader, 略してOCR)は郵便番号の自動読取区分装置に代表されるような、字種(英数字記号のみ、片仮名も含む、漢字・平仮名も含む等)、変形自由度(活字、手書き等)、筆記具(鉛筆、OCRボールペン、無制限等)、字体(単一、複数等)、文字ピッチ(固定、自由等)、搬送系(OCR用紙、上質紙、普通紙等)、識別水準(個別、単語、文章等)、フォーマット制御(固定、自由等)に制約がある文書に対してのものが実用化されてきた。しかし、ワードプロセッサ、パーソナルコンピュータの普及によって、多様な字種、字体、ピッチの文書の作成が容易になり、また、文書のフォーマットも複雑化、多様化してきたこたにより、制約の少ない(無い)OCRの開発が望まれてきた。

最近では、漢字認識、手書き文字認識の発展、半導体技術の進歩によって、これまで困難とされてきた複雑かつ高度な処理が可能なものになってきている。また、ワークステーション、パーソナルコンピュータの普及によって、膨大な量の文書(新聞、雑誌、書籍等)のデータベース化の社会的要望が高まっている。このため、より制約の少ないOCRの開発、実用化の努力がなされ、OA(オフィスオートメーション)、FA(ファクトリーオートメーション)、EA(エンジニアオートメーション)等の各分野で応用されてきている状況である。

このように、文字認識技術はOCRの開発、実用化という大きな成果を挙げている。しかし、現在のOCRはどのような文書でも認識できるわけではないし、文字認識能力も人間に比べるとまだ劣っている。よって、より完璧なOCRの開発、実用化が求められてい

る [1][2]。

1.2 本研究の目的

文字認識はいくつかのステップを通して行われるが、中でも重要であるのが文字切り出し技術である。本研究では、文字切り出しを高精度に行わせるための前処理として、英文字抽出法を提案する。

これまで、日本語文書用の文字切り出し法や英語文書用の文字切り出し法は数多く提案されているが、英文字の混在している日本語文書に対してはあまり提案されていない。なぜならば、英文字が混ざった日本文は英文字の不定なピッチと大きさによる切り出しの難しさと、日本語の分離文字による切り出しの難しさが互いに影響し合い、文字切り出しをより難しくしているからだ。

この問題を解消するための1つの方法として、今回、あらかじめ英文字部分を抽出することによって、従来提案されている日本語文書用の文字切り出しと英語文書用の文字切り出しを併用してより高精度な文字切り出しを実現しようというわけである。そして、文字認識全体としての認識率の向上を図るのである。

1.3 本論文の構成

本論文の構成は以下の通りである。

第1章 序論であり、本研究の背景と目的について述べる。

第2章 これまで提案されてきた文字切り出し法とその原理をまとめ、そこに含まれている問題点を明らかにする。

第3章 第2章で述べた問題点をふまえ、ここに新たに文字切り出しの前処理として、英文字抽出法を提案する。

第4章 第3章で提案した英文字抽出法の効果を認識実験を通して検討する。また、このために、第2章の問題点を解消するための簡易的な文字切り出し部、辞書を作成する。

第5章 本研究の結論、今後の課題を述べる。

第 2 章

文字切り出しとその問題点

2.1 はじめに

一般に文字認識は観測、前処理・正規化、特徴抽出、識別の 4 つ過程によって行われている。まず、観測とは、用紙から文書画像をスキャナ等で読み取ることである。次に、前処理・正規化とは、読み取った文書画像を文字単位に分割し、特徴抽出し易いように変形することである。そして、特徴抽出とは、認識のための情報を文字毎に抽出することである。最後に、識別とは、文字毎に得られた特徴と標準パターンを参照して各文字を断定することである。完全な文字認識のためには、これら 4 つの技術の発展はどれも必要不可決であるが、本研究では、中でも前処理部分に着目する。

前処理にもいくつかの段階 (スムージング、文字切り出し、正規化、細線化等) があるが、中でも文字切り出しは重要な技術と言える。なぜならば、文書画像から文字を正確に抽出できなければ、どんなに素晴らしい特徴抽出部、識別部があっても役に立たないものになってしまうからである。

これまで、いろいろな文字切り出し法が多くの研究者たちによって提案されてきているが、全ての文字に対して有効な方法はまだ発見されていない。ここでは、これまでの文字切り出し技術を振り返り、それらに残されている問題点をまとめることにより、新たな文字切り出し法の可能性を考える。

2.2 基本的な文字切り出し法

初期の OCR では文字の大きさ、配置が決まっている文書 (帳表等) を対象にしていたため、文字の切り出しの必要は無かった。しかし、OCR 技術の発展には、大きさ、配置等の制約の無い文書への応用が必要であり、それを実現するためには文字切り出し技術が重要となってきた。このため、様々な文字切り出し法がこれまで考案されてきている。

2.2.1 射影 (投影) 分布を用いる方法

射影 (投影) 分布とは、与えられた画像をある方向に投影した時にあらわれる黒画素の分布のことである。具体的には、投影方向に沿って計測した画素の度数分布のことである。

これを用いた切り出し方法 [3] は次のように行う。あらかじめ行単位に分離された画像に対して垂直方向に射影 (投影) 分布をとり、分布の山の部分に文字らしきものがあると判断し、分布の谷 (画素数 0) の部分を境界にして切り出していく。

この方法では、最初から垂直方向に各文字が分離していなければ、切り出しミスが生じてしまう。

図 2.1: 射影 (投影) 分布を用いた文字切り出しの例

2.2.2 連結領域に着目する方法

黒画素が連結している領域を文字らしいとして、連結領域毎に切り出していく方法 [4] である。連結の概念には 4 連結、8 連結 [5] があるが、文字の形状特徴 (縦線、横線以外に斜め線、曲線もありうる) を考慮して、主に 8 連結が用いられている。

この方法では、垂直方向に各文字が分離していなくても連結さえしていれば切り出しは成功する。しかし、文字自身が最初から分離している場合は切り出しミスになってしまう。

2.2.3 その他

前出の 2 つの方法が広く用いられているのであるが、その他にも切り出し方法は提案されている。例えば、統計的なモデルで文字ピッチを推定して、動的計画法を用いて切り出しを行う方法 [6] が考案されている。この方法は、固定ピッチの文書では効果的だが、欧文を含むような不定ピッチには適用できないという問題点がある。

2.3 基本的な文字切り出し法の問題点

日本語の特徴、英語の特徴等により、前節で紹介した切り出し方法では完全に個々の文字を切り出すことは難しい。以下に日本語、英語の各々の場合の問題点を挙げてみることにする。

2.3.1 文字切り出しにおける日本語の特徴上の問題点

日本語のみの文字切り出しには以下のような問題点がある。

1. 分離文字の存在。

図 2.2: 分離文字の例 (「に」が分離している)

2. つぶれ等による文字同士の接触。

図 2.3: 文字の接触の例 (「組織」が接触している)

2.3.2 文字切り出しにおける英語の特徴上の問題点

英文字の切り出しのみを考えると、以下のような問題点が挙げられる。

1. 不定な文字ピッチ (3.4.1 参照)

   l と m、i と w のように、文字間のピッチに大きな開きが見られる。

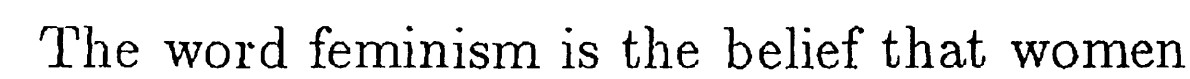


図 2.4: 英語の一般文書 (不定ピッチ) の例

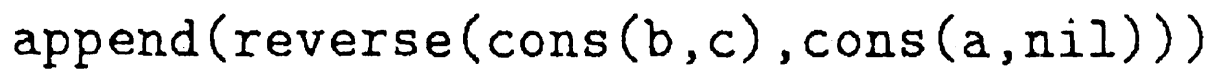


図 2.5: 英語の一般文書 (固定ピッチ) の例

2. 入り組み文字

- 印刷後の仕上がりを美しく見せるための文字間隔を詰める処理 (kerning)

図 2.6: kerning の例 (“aj” が入り組んでいる)

- 文字自体が斜めに傾いている斜体 (イタリック体等)

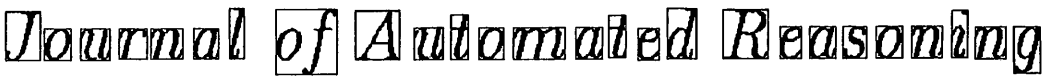


図 2.7: イタリック体の例 (“of” が入り組んでいる)

3. 文字間の接触

- 印刷の際、決まって 2 つ以上の文字が接触してしまう連字 (ligature)

図 2.8: ligature の例 (“ffi” は f と f と i が接触してできた連字)

- 人間が書く筆記体のようにわざと接触させて書れた文字
- コピーなどの結果、つぶれが生じてしまい、接触してしまった文字

2.3.3　対策

このように、日本語には日本語の、英語には英語の切り出しにおける問題点がある。よって、日本語 OCR では日本語を正確に切り出すための処理 (ピッチ情報を用いた分離文字の統合等) を、英語 OCR では英語を正確に切り出すための処理 (接触文字の分離 [7]、文字単位の傾き補正 [7] 等) をしている。

2.4 日本語と英語が混在している文書の切り出し法

一般に日本語文書は漢字、平仮名、片仮名の他に英数字を用いていることが多く、英数字の大きさ、配置の不規則さによる生じる問題と日本語の分離文字によって生じる問題とが互いに影響し合い、文字の切り出しを更に難しくしている。

これまで、この問題を解決するため、ピッチ情報を用いて切り出す方法 [8]、文字列のレイアウト上の特徴を用いて切り出す方法 [9]、容易に切り出せる文字から切り出していく方法 [10] 等、様々な手法が提案されてきている。

2.5 まとめ

このように、現在の文字切り出し技術では、日本語のみの文書に対して、英語のみの文書に対して、日本語と英語が混在していても定ピッチな文書に対しての3つの場合はほぼ解決したと言える。しかし、日本語と英語が混在した不定ピッチの文書に対する効果的な切り出し法はあまり考案されていない。よって、本研究では、日本語と英語が混在した文書の文字切り出しのために、文書全体から英語の領域を抽出することを目標とする。そうすることによって、英語の領域は英語用の切り出しを、その他の領域は日本語の切り出しをすることができるようになり、全体的な切り出し成功率が高まるだろうと考えられる。英語の抽出法は次章で提案する。

第 3 章

日本語文中からの英文字抽出

3.1 はじめに

日本語文書の中には、種類によって英文字が混ざっていることがある。例えば、科学技術関連の文書 (論文等) がその代表である。日本語と混ざっている英文字は文章内の文字ピッチを乱すため、文字の切り出しが難しくなり、その結果、認識率を低下させる原因となっている。

この章ではこの文字ピッチの乱れを逆に利用して、文章中から英文字部分の抽出をする。図 3.2 のフローチャートは英文字抽出の一連の流れを示したものである。次節からは、それぞれの段階ごとに説明していくことにする。

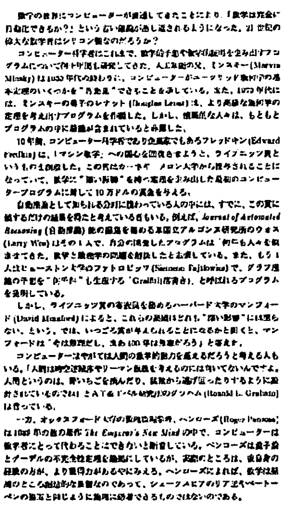


図 3.1: 英文字が混在する日本語文書例

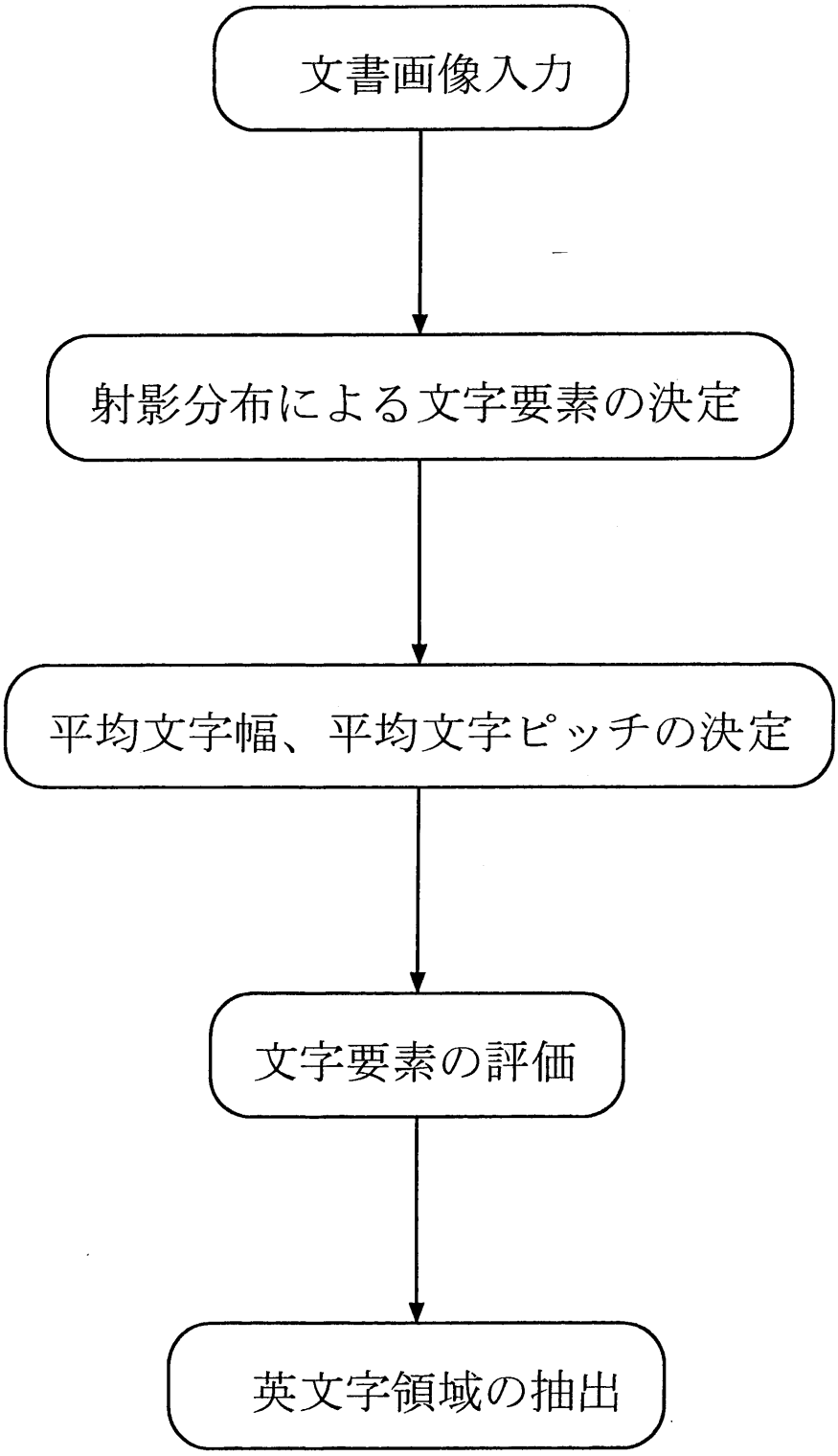


図 3.2: 英文字抽出の概略

3.2　文書画像入力

認識対象となる文書は 1 ページずつスキャナで読み込む。この時の読み取り密度は 400dpi である。そうして得られた文書画像に対して、次節以降の処理を施すのである。

本研究では横書きの日本語文書 (特に英文字を含んだ) を中心に扱っていく。

3.3　射影分布による文字要素の決定

本研究では文書画像から行を抽出し、その行から各文字要素を決定するために射影分布 (2.2.1 参照) を利用している。以下にその方法を簡単に説明する。

まず、1 ページ分の文書画像が与えられたならば、その文書に対して水平方向に射影分布をとる。そして、求めた射影分布の山の部分に行らしきものがあると判断して、射影分布の谷の部分 (画素数 0) を境界として行を抽出するのである (図 3.3 上)。

次に行の画像が得られたならば、各行ごとに文書に対して垂直方向に射影分布をとる (図 3.3 中)。そして、行抽出と同様に射影分布の山の部分を文字らしきものがあると判断し、それを文字要素として登録する (図 3.3 下)。ここで言う登録とは、その文字要素の番号 $i \in [0, num)$(その行に i 番目に登場) を与え、その高さ height(i) と横幅 width(i)、そして、横方向の座標 s-width(i) の各データを保存することである。これらのデータは次節以降に用いることになる。

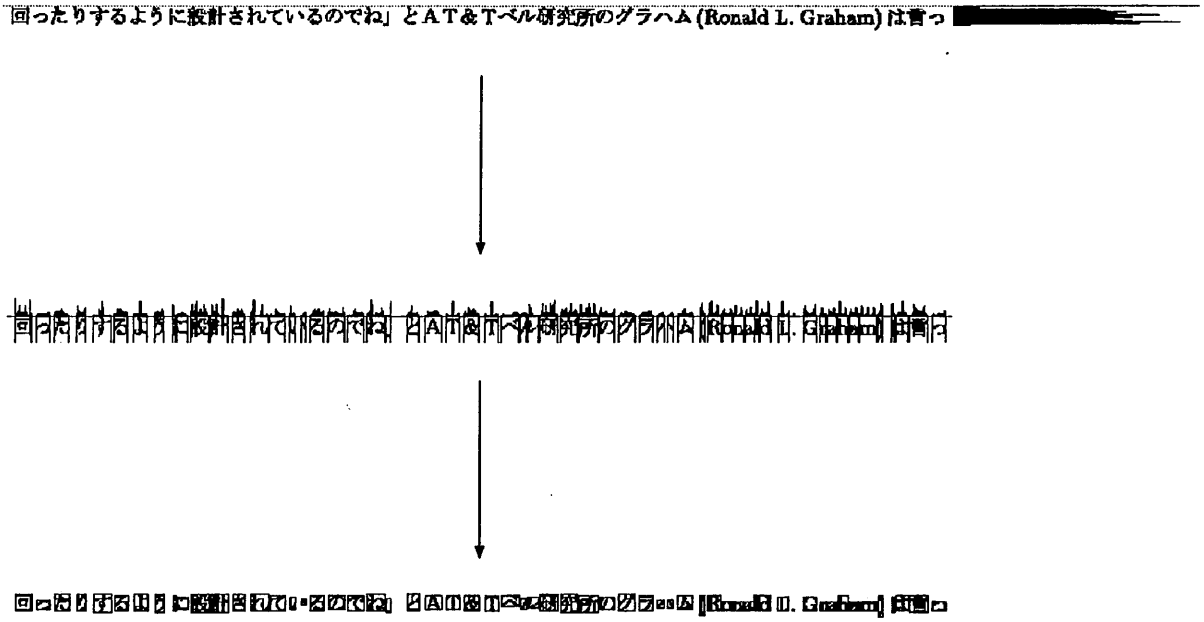


図 3.3: 文字要素決定までの過程

3.4　平均文字幅、平均ピッチの決定

日本語文書を行単位で見ると、日本語(漢字、平仮名、片仮名、記号類)のみしかあらわれないものと日本語に加えて英文字もあらわれるものと2種類に分類できる。(英文字と記号類のみしかあらわれないものがある場合も考えられるが、それは稀であるので分類には含んでいない)。それらの中で日本語と英語との間の何らかの違いを見出すことができれば英文字抽出も可能になってくる。この節では、その違いを文字の外見上の特徴のみ(本研究では文字幅と文字ピッチ)で発見し、それを英文字抽出の方法に取り入れることを考える。

3.4.1　文字幅と文字ピッチ

文字幅とは単純に文字の横幅のことである。ただし、本研究ではこの時点で完全に個々の文字を切り出してはいないので正確には文字要素の横幅を指している。また、文字ピッチとは隣合う文字(要素)とその文字(要素)自身との間にできる空白の半分をその文字(要素)の横幅に加えたものである。これらを式であらわすと以下のようになる。

$$\text{文字幅} \ : \ \mathrm{Width}(i) = \mathrm{width}(i)$$

$$\text{文字ピッチ} \ : \ \mathrm{Pitch}(i) = \mathrm{width}(i) + \mathrm{space}(i) + \mathrm{space}(i+1)$$

但し、

$$\mathrm{space}(i) = \begin{cases} \frac{\text{s-width}(i+1) - \text{s-width}(i) - \text{width}(i)}{2} & (0 < i < num) \\ 0 & (i = 0, i = num) \end{cases}$$

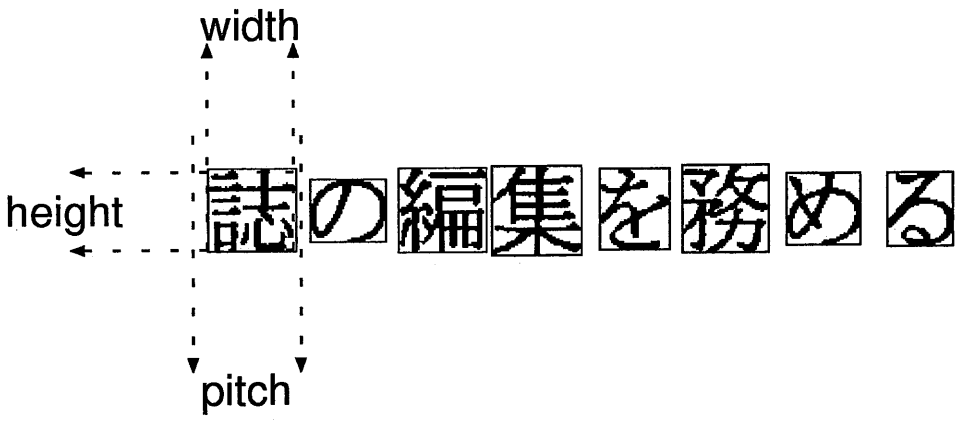


図 3.4: 文字要素の外見情報の定義

3.4.2 一行あたりの文字幅、文字ピッチの分布

日本語のみしかあらわれない行、英語のみしかあらわれない行、両方ともあらわれる行の3種類についてそれぞれの文字幅、文字ピッチの分布を調査する。これは、一行だけの情報からその行を構成している各文字要素を英文字と非英文字とに分類するための尺度を見出すことを目的とする。一般に、日本語は文字の高さと横幅がほぼ同じであるのに対して英文字はそれが成り立たないので、それを実験的に確かめるのである。

3.4.2.1 実験方法

日本語のみの行イメージと英語のみの行イメージと両方混在する行イメージを5行ずつ用意して、それぞれについてその行を構成している文字要素の文字幅と文字ピッチを調べる。そして、得られた結果を基に各種の行について、文字幅と文字ピッチのヒストグラムを作成し、日本語と英語のそれぞれの特徴を調べる。ここで、ヒストグラムは単純に得られたデータの度数を計測するのではなく、得られたデータを長さ5文字要素の方形フィルタを通して平滑化して計測することにする。

3.4.2.2 調査対象

日本語のみの行イメージ 英文字を一切含まないテキスト (社説) から引用。

英語のみの行イメージ ローマン体中心 (一部イタリック体も含む) の日本語を一切含まないテキスト (英語のエッセイ) から引用。

両方を含む行イメージ 英文字を平均20%含んだテキスト (論文誌、科学雑誌) から引用。

3種類の行イメージとも、LaTeX で10pt の大きさに修正した文書をスキャナから400dpi の解像度で取り込み、得られた文書画像に前節で説明した行抽出を適用して抽出したものである。

3.4.2.3 結果

調査結果はグラフに示してある。各々のヒストグラムから以下のことが読み取ることができる。

日本語のみの行イメージ 文字幅は、その行の高さの60%以上のところに度数が集中し、文字ピッチは、その行の高さの80%以上のところに度数が集中している (図3.5)。

英語のみの行イメージ 文字幅、文字ピッチの両方とも、その行の高さの50%付近に度数が集中している (図3.6)。

日本語と英語が混在する行イメージ 日本語のみの行の結果と英語のみの行の結果が混ざり合ったような分布をしている。つまり、日本語の度数のピークと英語の度数のピークが見られるのである (図3.7)。

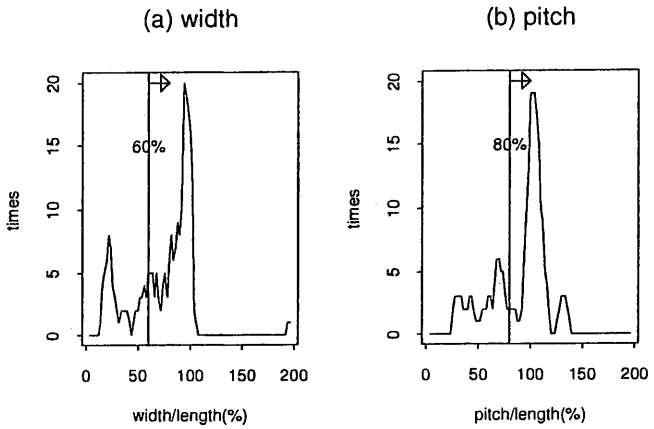


図 3.5: 日本語のみの行イメージの文字幅、文字ピッチの度数分布例

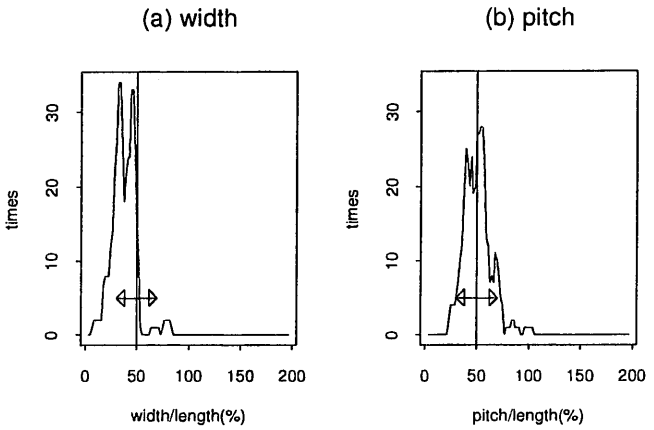


図 3.6: 英語のみの行イメージの文字幅、文字ピッチの度数分布例

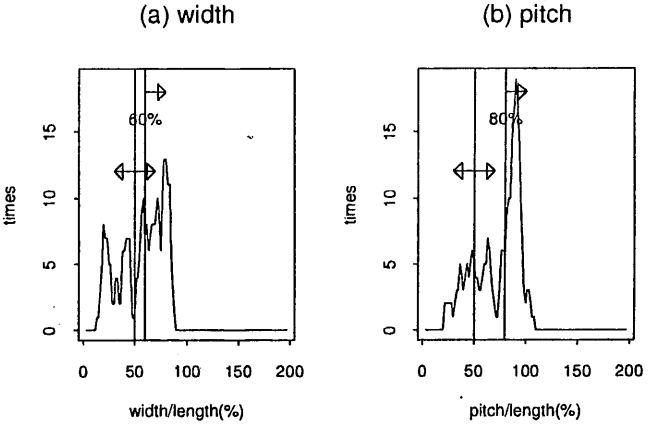


図 3.7: 日本語と英語が混在している行イメージの文字幅、文字ピッチの度数分布例

3.4.3 平均文字幅、平均文字ピッチの推定

ここで言う平均文字幅、平均文字ピッチとは与えられた行イメージの文字要素全部についての平均ではなく、その行に含まれている文字要素が全て日本語と仮定した時の平均である。これらは、前節で調査した結果に基づいてそれらを定めることになる。

まず、平均文字幅はその行の文字要素から文字幅のヒストグラムをとり、長さ5文字要素のフィルタを通して平滑化する。そうして得られたデータの中でその行の高さの60%以上の文字幅の度数を調べ、最大値(複数の場合、最も大きい文字幅)をその行の平均文字幅とする。

次に、与えられた行の全文字要素の文字ピッチのヒストグラムをとり、平均文字幅を求める時と同様に、長さ5文字要素のフィルタに通して平滑化する。そうして得られたデータの中でその行の高さの80%以上の文字ピッチの度数を調べ、最大値(複数の場合、最大ピッチ)をその行の平均文字ピッチとする。

以上の2つの平均データは日本語らしさの尺度として、次節以降に用いることになる。

3.5 文字要素の評価

前節で求めた平均文字幅、平均文字ピッチを用いて各文字要素の日本語らしさ、英語らしさをあらわすスコアを決める。

3.5.1 スコア

各文字要素の文字幅と文字ピッチを平均文字幅、平均文字ピッチと比較してスコアをつける。スコアリングは文字幅、文字ピッチとも以下に述べる方法で行う。但し、平均文字幅を Width(ave)、平均文字ピッチを Pitch(ave) とする。

1. 平均値がその文字要素の値以上、かつ、隣の文字要素との間のスペースを文字幅に加えた値未満ならばその文字要素の値が平均にほぼ一致するので、日本語らしいとみなしてスコアは1となる。

$$\mathrm{Width}(i) \leq \mathrm{Width}(ave) < \mathrm{Width}(i) + 2 \times \mathrm{space}(i) \longrightarrow \mathrm{widthscore}(i) = 1$$

$$\mathrm{Pitch}(i) \leq \mathrm{Pitch}(ave) < \mathrm{Pitch}(i) + \mathrm{space}(i) \longrightarrow \mathrm{pitchscore}(i) = 1$$

図3.8の1.がこの場合の具体例である。「連」の文字幅(矩形の横幅)が平均文字幅と同じ位の長さなので、スコアは1となる。

2. 平均値が2つ以上の文字要素の値とそれらの間のスペースを加えた値以上、かつ、更に隣の文字要素との間のスペースを加えた値未満ならば、2つ以上の文字要素を合併

すれば日本語らしいと判断できるが、1 よりは日本語らしさが低いとみなし、スコアは 0 となる。

$$\sum_{j=0}^{n}\mathrm{Width}(i+j)+2\times\sum_{k=0}^{n-1}\mathrm{space}(i+k) \leq \mathrm{Width}(ave) < \sum_{j=0}^{n}\{\mathrm{Width}(i+j)+2\times\mathrm{space}(i+j)\}$$

$$\longrightarrow \mathrm{widthscore}(i) = 0 \text{ (但し } n > 0)$$

$$\sum_{j=0}^{n}\mathrm{Pitch}(i+j)+\sum_{k=0}^{n-1}\mathrm{space}(i+k) \leq \mathrm{Pitch}(ave) < \sum_{j=0}^{n}\{\mathrm{Pitch}(i+k)+\mathrm{space}(i+j)\}$$

$$\longrightarrow \mathrm{pitchscore}(i) = 0 \text{ (但し } n > 0)$$

図 3.8 の 2. がこの場合の具体例である。「に」は 2 つの文字要素 (矩形で囲まれたもの) で構成されているが、それらを合わせた文字幅が平均文字幅と同程度であるため、スコアは 0 となる。

3. 平均値がその文字要素の値以下ならば、その文字要素は幾つかの文字 (要素) が接触しているものと判断できるが、日本語が接触しているのか英語が接触しているのかまでは分からないのでスコアは 0 としておく。

$$\mathrm{Width}(ave) < \mathrm{Width}(i) \longrightarrow \mathrm{widthscore}(i) = 0$$

$$\mathrm{Pitch}(ave) < \mathrm{Pitch}(i) \longrightarrow \mathrm{pitchscore}(i) = 0$$

図 3.8 の 3. がこの場合の具体例である。「組織」が 1 つの文字要素となっているため、その文字幅が平均文字幅を越えてしまっているので、スコアは 0 となる。

4. 平均値がその (それらの) 文字要素の値 (とそれらの間のスペースを加えた値) 以上、かつ、その隣の文字要素との間のスペースを加えた値以上ならば、平均値に遠い値を持っているので、英語らしいと判断され、スコアは-1 とする。

$$\sum_{j=0}^{n-1}\{\mathrm{Width}(i+j)+2\times\mathrm{space}(i+j)\} < \mathrm{Width}(ave) < \sum_{j=0}^{n}\mathrm{Width}(i+j)+2\times\sum_{k=0}^{n-1}\mathrm{space}(i+k)$$

$$\longrightarrow \mathrm{widthscore}(i) = -1 \text{ (但し } n > 0)$$

$$\sum_{j=0}^{n-1}\{\mathrm{Pitch}(i+j)+\mathrm{space}(i+j)\} < \mathrm{Pitch}(ave) < \sum_{j=0}^{n}\mathrm{Pitch}(i+j)+\sum_{k=0}^{n-1}\mathrm{space}(i+k)$$

$$\longrightarrow \mathrm{pitchscore}(i) = -1 \text{ (但し } n > 0)$$

図 3.8 の 4. がこの場合の具体例となる。「S」はそれ自身の文字幅は平均文字幅に比べて小さすぎるし、隣合う他の文字要素と組み合わせてもその文字幅は平均文字幅と同程度にはならないので、スコアは−1 となる。

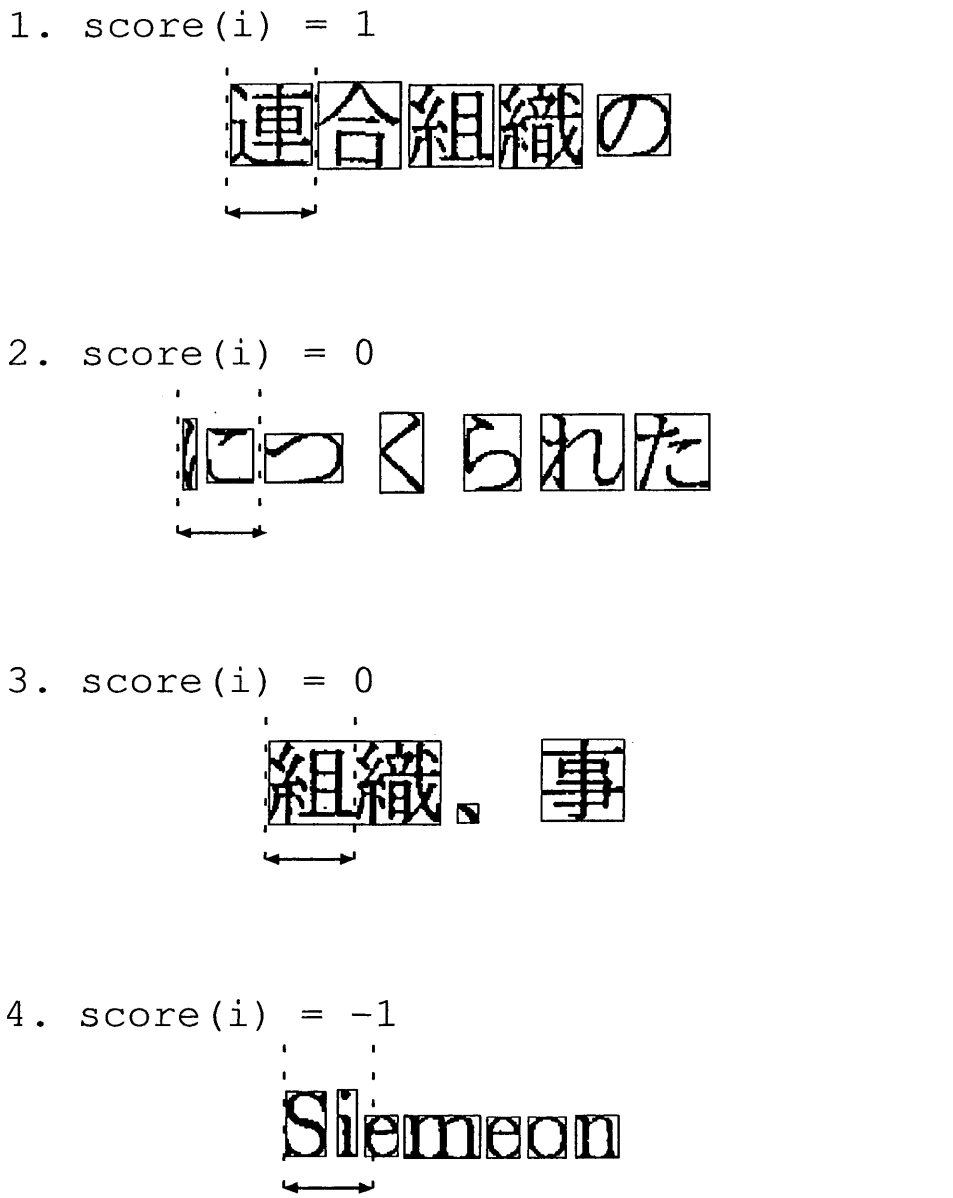


図 3.8: 文字要素自身のスコアリングの具体例

以上のようにスコアリングしたならば、文字幅によるスコア widthscore(i) と文字ピッチによるスコア pitchscore(i) を加えて、その文字要素自身のスコア score(i) とする。

$$\text{score}(i) = \text{widthscore}(i) + \text{pitchscore}(i)$$

ここで、文字幅と文字ピッチの両方の情報を用いる理由は、文字幅だけであれば、平均より文字幅が狭い日本語が英語らしいと判断されるからであり、文字ピッチだけであれば句読点などの記号を英語らしいと判断してしまうからである。

3.5.2　スコアのフィルタリング

前節でも述べたように、本研究では文字要素の英文字らしさの度合として、文字幅と文字ピッチの両方によるスコアリングを行った。しかし、文字要素自身のスコアのみでそれが英文字かどうかを判定するには条件が少なすぎると思われる。そこで、周囲の文字要素の影響を考慮するために、長さ 5 文字要素の方形フィルタに各文字要素のスコア score(i) を通して、その結果を最終的な文字要素のスコア Score(i) とする。

$$\text{Score}(i) = \sum_{k=-2}^{2} \text{score}(i+k)$$

但し、 $\text{score}(-2) = \text{score}(-1) = \text{score}(num) = \text{score}(num+1) = 0$ とする。

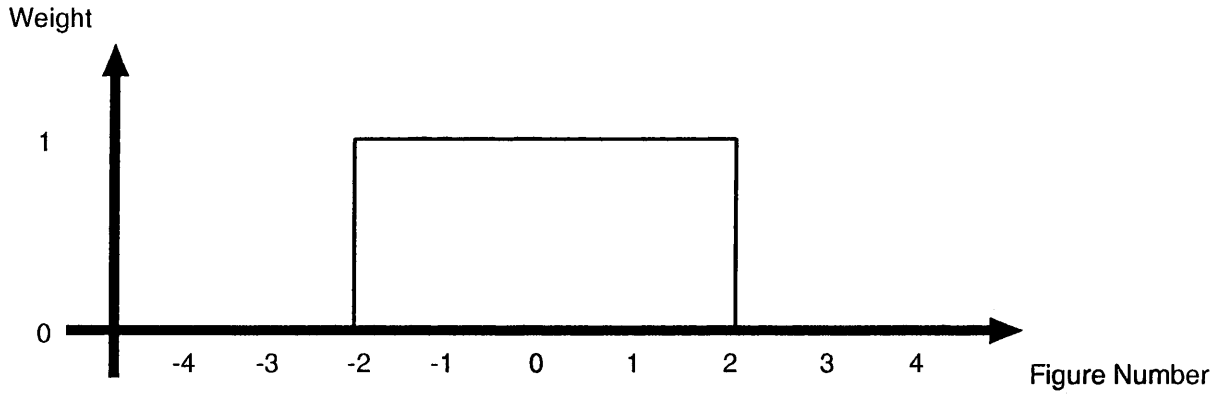


図 3.9: 長さ 5 文字要素の方形フィルタ

3.6 英文字領域の抽出

3.6.1 英文字領域の判定

前節までに求めた文字要素の最終的なスコア Score(i) を基にして、各文字要素が英文字であるかどうかの判定を行う。

まず、最終的なスコア Score(i) の上限、下限を考える。文字要素自身のスコア score(i) がとりうる範囲は$-2 \leq \text{score}(i) \leq 2$ であるから、5 文字要素分のスコアを加えた最終的なスコア Score(i) がとりうる範囲は$-10 \leq \text{Score}(i) \leq 10$ となる。

次に、その範囲内で英語と日本語がちょうど分離するポイント (閾値) を決定する。スコアリングの性質から考えると、$\text{Score}(i) < 0$ の範囲が英語らしいと言える。だが、最終的にはスコアにフィルタをかけてしまうため、文字要素自身のスコア score(i) が 2 点で日本語と判断できても、周囲の文字要素が全て-2 点で英語と判断されていれば、方形フィルタによって最終的なスコア Score(i) は-6 点になってしまい、日本語を英語と誤って判定する可能性が生じてしまう。そこで、実験によって、閾値を-6 から 0 に変化させた時に、誤抽出がどの程度発生するかを確かめることにして、その結果により、英文字判定の閾値を定めようと思う。

3.6.2 英文字らしさの閾値の決定

3.6.2.1 決定方法

英文字領域の判定に用いるスコアの閾値を実験的に求める。方法としては、任意に選んだ英語混じりの日本文の行イメージをここまで説明してきた手法を用いて、閾値を-6 から 0 に変化させて英文字抽出を行い、それぞれの場合の誤抽出率を計算する。

抽出対象 LaTeX で作成した 10pt の文書をスキャナで 400dpi の解像度で読み込んだイメージを 12 行分使用した。文書の内容は、英語を約 50%含んだ日本語文書 (論文誌、科学雑誌) からの引用である。

3.6.2.2 結果

表 3.1: いろいろな閾値での英文字領域の抽出率

閾値	誤抽出率	抽出率
0	8.77%	86.83%
-2	4.09%	72.41%
-4	1.46%	59.25%
-6	0.29%	36.36%

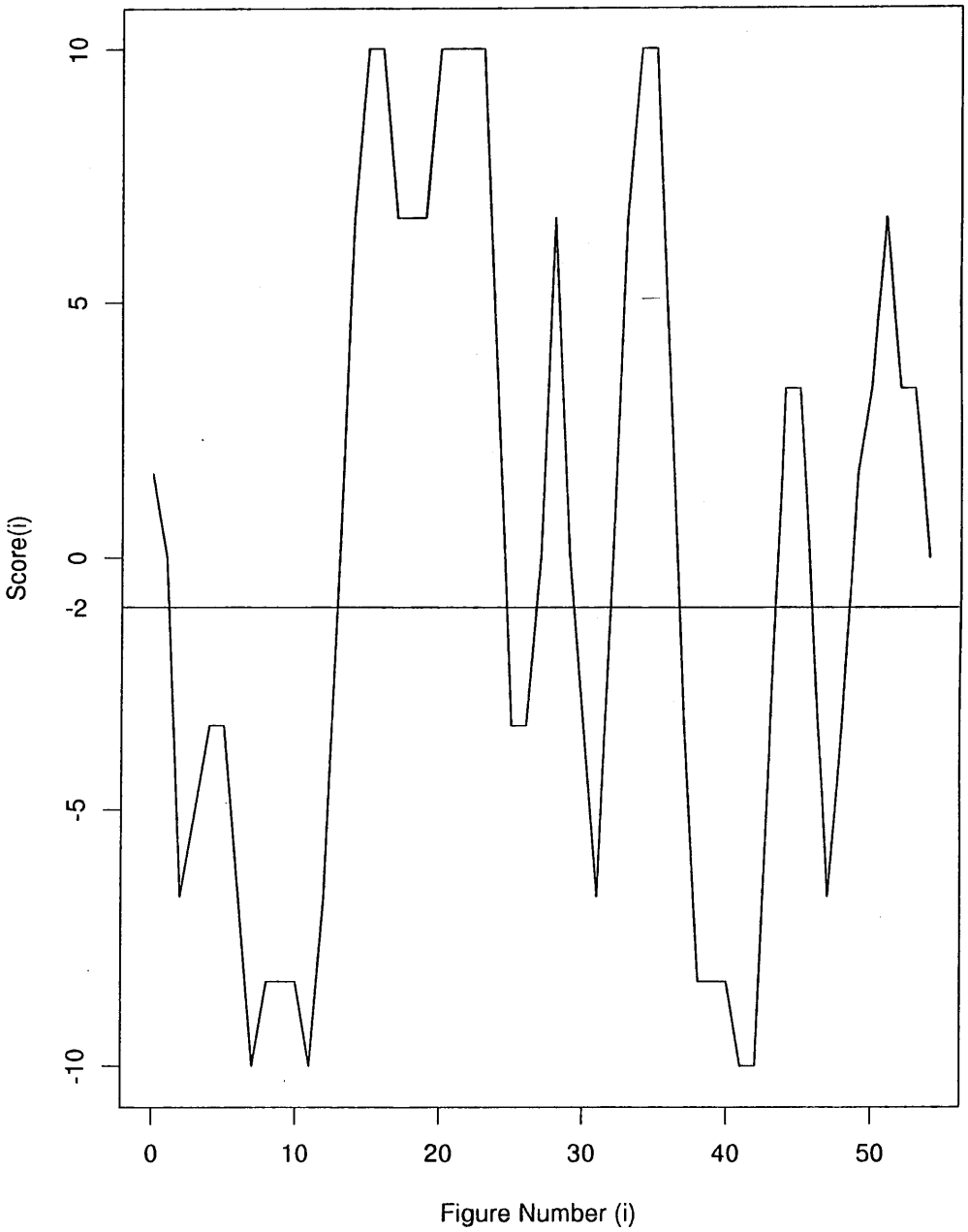


図 3.10: スコアリングの様子

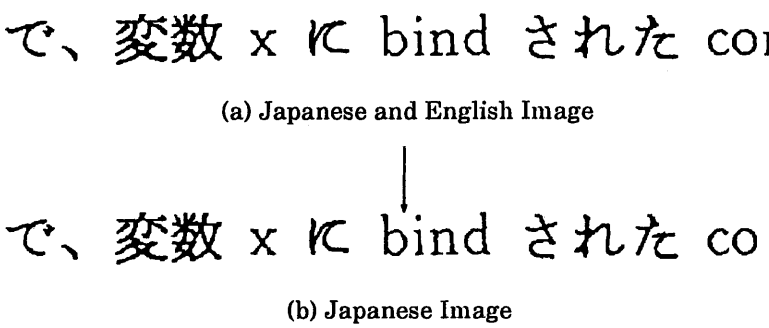
表 3.1 が予備実験の結果である。この結果より、閾値を高く設定すると英文字の抽出率は高まるが、それと同時に日本語を誤抽出する確率も高くなってしまうことが分かる。だからと言って、閾値を低く設定すると日本語を誤抽出する確率は低くできるが、抽出できる英文字数も少なくなってしまい、英文字抽出の意味が無くなってしまう。以上を考慮して、今回は抽出率が比較的高く、誤抽出率も低い-2 を閾値と定めることにする。

3.6.3　英文字抽出結果の検討

ここまで説明してきた英文字抽出法による英文字の抽出状況を検討する。実験は閾値を決める段階で既に行ってあるのでそこで得られた結果を検討することになる。

3.6.3.1　英文字の抽出率の検討

英文字の抽出率は、72.41%とやや低い結果であった。よって、このような結果となった原因を実際の抽出例を見ながら考えることにする。



(a) Japanese and English Image

(b) Japanese Image

図 3.11: 抽出できなかった英文字の例

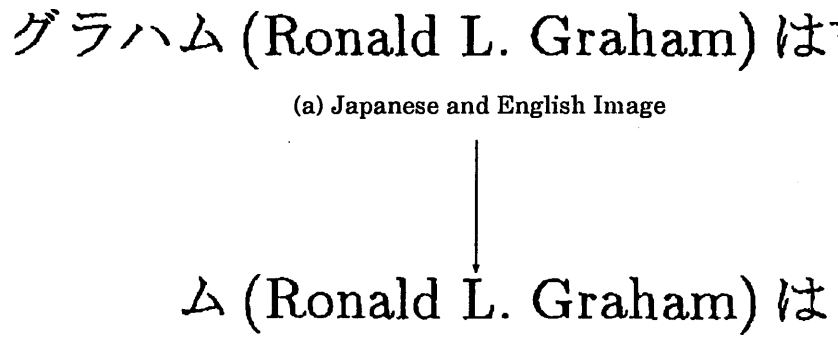
抽出できなかった英文字についてまず言えることとしては、字数の短い単語 (3 文字前後) が多いことが挙げられる。文字この原因としては、最終的なスコア Score(i) を決定する時の用いる方形フィルタにあると考えられる。つまり、フィルタによって、本来その文字要素自身が持っていた英文字らしいという情報 (スコア) が周囲の日本語のスコアに打ち消されてしまったためであろうと考えられる。具体的な例を図 3.11 に示した。この図の中の「x」という文字は、両隣とも日本語であることにより、それ自身の英文字らしさのスコアガ打ち消さたために英文字として抽出できなかったのである。

もう 1 つ言えることは、英文字同士が接触したものの未抽出が挙げられる。これは、接触のために、その文字幅 (文字ピッチ) が平均文字幅 (平均文字ピッチ) に近くなってしまい、日本語と誤って抽出されたのだと考えられる。これの具体的な例も図 3.11 に示してある。この図の中の「bind」という英単語は画像上では接触していないが、「bi」と「nd」が

それぞれ平均文字幅に近くなってしまったために、その文字要素自身に英文字らしさのスコアを与えることができず、その結果、抽出できなかったのである。

3.6.3.2　日本語の誤抽出の検討

次に、日本語の誤抽出率について考える。数値的には4.09%とやや低めであるが、ある目立った誤抽出例が見られた。それは、英文字領域のちょうど境界あたりに日本語が残留してしまっている例である。これも方形フィルタの影響であると考えられる。これの具体例は図3.12に示してある。英文字領域の境界部分に日本語の「ム」と「は」が残留している様子が分かる。これは、完全に方形フィルタによる悪影響が原因であると考えられる。



(a) Japanese and English Image

(b) English Image

図 3.12: 日本語の誤抽出の例

3.6.3.3　抽出成功例の分析

最後に、抽出成功例を見ることにする。特に目立ったのは、英文字が連続して多く並んでいる場合で、良好に英文字抽出が行われていることが分かった。このことは、スコアにフィルタをかけたことによる良い影響と言えよう。

3.7　まとめ

本章では、文字切り出しの前処理として、英文字抽出法を提案した。そして、文書画像入力、射影分布による文字要素の決定、平均文字幅・平均文字ピッチの決定、文字要素の評価、英文字領域の抽出という5段階に分かけて本手法の詳細を説明した。また、この英文字抽出によりどれだけの英文字を抽出できるのかを英文字判定のための閾値を決める作業と同時に実験した。その結果、文書中の英文字の約75%を抽出することに成功した。しかし、それと同時に、方形フィルタをかけることが英文字抽出の失敗例を引き起こしているという問題点も明らかにされた。

このように、英文字抽出はまだ未完成といっても過言では無いが、ある程度の良好な結果も得られたのせ、次章では、実際の認識システムに英文字抽出を組み込んだ時にどれだけ認識に対して良い効果が得られるかを検討しようと思う。

第 4 章

英文字抽出の有効性の検証

4.1 はじめに

前章で、今回提案する英文字抽出の方法を説明した。しかし、実際に文字認識の性能にどのように影響するかは、英文字抽出の結果のみでは判断できない。よって、この英文字抽出を実際に文字認識システムに組み込んで、英文字抽出の有効性を検証しようと思う。

実際にどのように英文字抽出を組み込むかは次のとおりである。

1. スキャナで読み込んだ文書画像を前章で説明した英文字抽出法を用いて、英文字領域とその他の領域に分割する。

2. 英文字領域には英語のための文字切り出しをして、その他の領域には日本語のための文字切り出しをする。

3. 英文字領域の切り出し結果を英語専用辞書を用いて識別する。また、その他の領域の切り出し結果には英語も含んだ日本語用辞書を用いて識別する。

ここで用いる日本語のための切り出し及び日本語用辞書は本研究室で開発された高速高精度文字認識システム SEIUN[11] のためのものである。今回はそれをシミュレートしたプログラムを用いて実験をしている。一方、英語領域の認識のための切り出しと辞書は今回この検証用として簡易的なものを作成した。これらについては次節以降で説明していく。

4.2 英文字領域用の切り出し

第 2 章で述べたように、英語には英語独自の特徴があり、これを生かした文字切り出しを行わなければならない。ここでは英語の特徴上の問題点である斜体の切り出しも一部考慮した文字切り出し法を作成した。

4.2.1　切り出しの手順

英文字領域の切り出し手順は以下の通りである。

1. 英文字領域に文書に対して縦方向の射影分布をとり、それを基に各文字要素へ分割する。

2. 得られた各文字要素について、隣合う文字要素との距離を調べて、1 画素分の距離しかないものについてのみ、それらの文字要素を合併して 1 つの文字要素とする。

3. 合併処理後、各文字要素の矩形比 Width(i)/ height(i) を求め、それが 1.3 以上のものについては分離処理を施す。

ここで、1. の射影分布による切り出しを中心にしたのは、英文字には日本語と違い、分離文字が存在しないからと、分布を求める処理が比較的単純であったからである。2. の合併処理の導入の理由は、かすれ等による文字の分離を考慮したためである。そして、3. の分離処理は、斜体対策に導入したものであり、詳細は次に譲る。ここで、分離処理へ送る文字要素の矩形比の閾値を 1.3 にしたのは、1 文字の矩形比の上限だろうと予測してのことであり、特に、実験等で詳しい検討は行っていない。

4.2.2　分離処理

イタリック体等の斜体が入り組んでしまった時、垂直方向の射影分布からのみでは切り出す位置を決定することができない。そこで、分離処理として、垂直方向に対して、5 °、10 °、15 °、20 °、30 ° の 5 種類の射影分布をとって、その中から最適な切り出し角度と切り出し位置を求める。実際には、以下のように行われる。

1. 分離処理に送られてきた文字要素について、垂直方向を基準にして、0 °、5 °、10 °、15 °、20 °、30 ° の各々の方向の射影分布をとる。

2. 各方向の射影分布の両端 10% を除いた部分の分布を調べて、最も小さい度数を持つ角度を切り出し角度として、その度数の最小値で斜めに切り出す。

3. 2. で最も小さい度数を持つ角度が複数あった場合、最も小さい度数が多くあらわれる角度を切り出し角度として、その度数の最小値で斜めに切り出す。

4. 3. で最も小さい度数の個数も一緒の場合、より垂直に近い角度を切り出し角度として、その度数の最小値で斜めに切り出す。

ここで、1. で垂直方向の射影分布もとる理由は、合併処理で誤って合併してしまったものを再び分離できるようにするためである。また、その他の 5 つの角度の設定には特別に理論的、実験的な根拠は無く、大体これくらいで切り出せるだろうと考えられる角度を任意に選んだだけである。また、これら 5 つの角度は厳密には正確では無く、画像の濃度を

表 4.1: 斜め切り出しの角度設定とその実際の角度

設定角度	shift 周期	実際の角度 ($\tan^{-1}\frac{1}{\text{shift}}$)
5 °	12	4.764 °
10 °	6	9.462 °
15 °	4	14.036 °
20 °	3	18.435 °
30 °	2	26.565 °

考慮して、最もそれらの角度に近いものを近似的に利用しているのである。表 4.1 はこれら 5 つの角度の設定状況をまとめたものである。

また、2. で射影分布を中央部を中心に見るのは、切り出しによって文字要素が分離するとすれば、その文字要素の中心付近にその境界がある可能性が高いだろうと考えられるからである。更に、両端に生成される極端に小さい領域が抽出されてしまうのを防ぐためでもある。

4.2.3　本切り出し法の評価

4.2.3.1　目的

ここで作成した切り出し法が日本語用の切り出し法 (分離文字の統合を含む) に比べて有効的であることを示さなければならない。なぜならば、ここで作成した切り出し法が日本語用の切り出しより劣っていたならば、前章で提案した英文字抽出の意味が無くなってしまうからである。

以上の理由から、本切り出し法と日本語用切り出し法との性能の比較実験を行う。

4.2.3.2　比較方法

日本語用切り出し法と本切り出し法を使って同じ英語の文書の切り出しをさせて、その切り出し成功率を調べる。

切り出し対象　ここで用いた英語のテキストは、LaTeX で 10pt の大きさに作成したもので、その内容は、英語のエッセイからの引用である。また、切り出し対象文字数は 1985 文字である。

4.2.3.3　結果

以下に結果を表で示す。

表 4.2: 切り出し法の性能比較結果

方法	切り出し率
本切り出し法	98.24%
日本語用切り出し法	98.09%

このように、全体的には改悪点 (分離処理による文字要素の過剰な分割) よりも改善点 (分離処理による入り組み文字の分離) の方が多く、本切り出し法でも効果的に英文字を切り出せることがわかった。

4.3　英文字専用辞書の作成

本研究では、英文字抽出によって、あらかじめ英文字領域が分かっている状態であるから、日本語を含まない、英文字だけの辞書で識別を行うことによって、より高精度で高速な認識が実現できると思われる。よって、次に述べるような英文字用簡易辞書を作成する。

4.3.1　英文字専用辞書の構成

今回作成した辞書の学習サンプルは、表 4.3 にある英文字記号 94 種を図 4.1 にある各フォントにつき 1 セットずつ計 3102 文字を用いた簡易的なものである。特徴量としては方向線素特徴量 [12] を用いている。

また、英語の特徴上の問題点であった連字 (ligature) も今回作成した辞書に取り入れた。具体的には、学習サンプルとして、TEX のフォント 4 種類 (Roman、**Bold**、*Italic*、*Slanted*) について表 4.4 の 5 種類の連字 1 セットずつ用いた。

表 4.3: 英文字記号 1 セット一覧

!	”	#	$	%	&	’	(	)	*	+	,	−	.	/					
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	@			
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z					[	\	]	^	_	‘				
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
u	v	w	x	y	z					{	|	}	~						

Times-Roman

Times-Italic

Times-Bold

Times-BoldItalic

AvantGarde-Book

AvantGarde-BookOblique

AvantGarde-Demi

AvantGarde-DemiOblique

Bookman-Light

Bookman-LightItalic

Bookman-Demi

Bookman-DemiItalic

Courier

Courier-Oblique

Courier-Bold

Courier-BoldOblique

Helvetica

Helvetica-Oblique

Helvetica-Bold

Helvetica-BoldOblique

Helvetica-Narrow

Helvetica-Narrow-Oblique

Helvetica-Narrow-Bold

Helvetica-Narrow-BoldOblique

NewCenturySchlbk-Roman

NewCenturySchlbk-Italic

NewCenturySchlbk-Bold

NewCenturySchlbk-BoldItalic

Palatino-Roman

Palatino-Italic

Palatino-Bold

Palatino-BoldItalic

ZapfChancery-MediumItalic

図 4.1: 辞書の学習サンプルに用いたフォント

表 4.4: 本辞書に組み入れた連字の一覧

連字	連字の構成
ff	f と f
fi	f と i
fl	f と l
ffi	f と f と i
ffl	f と f と l

4.3.2 英文字専用辞書の評価

4.3.2.1 目的

作成した英文字専用辞書が英文字領域の認識に効果的であるかどうかを日本語の辞書 (英語も含む) と比較実験をしてみる。

4.3.2.2 実験方法

英語専用辞書、日本語の辞書のそれぞれを用いて、同じ認識対象、同じ特徴量 (方向線素特徴量)、同じ識別法 (シティーブロック距離 [5] を用いたパターンマッチング法) で認識実験の比較をする。

認識対象 TEX のフォント 6 種類 (Roman、**Bold**、*Italic*、*Slanted*、Sans Serif、Typewriter) について 10pt のあらかじめ切り出されている英文字記号各 10 セットと TEX の 4 フォント (Roman、**Bold**、*Italic*、*Slanted*) で同じく 10pt の大きさの連字も各 10 セット用意した。

4.3.2.3 結果

以下の評価実験の結果を表にまとめておく。

表 4.5: 英文字専用辞書の性能比較結果

辞書	認識率
英文字専用辞書	80.14%
日本語の辞書	68.32%

このように、日本語の辞書での認識よりも英文字専用辞書での認識の方がかなり効果的であることが分かった。特に、斜体の認識率で英文字専用辞書と日本語用の辞書との間で大きな差が生じた。これは、日本語用辞書では斜体の学習データが不足していたのではないかと思われる。また、全体的な認識率の差の原因としては、日本語用辞書では、識別候補が英文字以外にも多く存在し、誤認識の可能性が高まったのではないかと思われる。

4.4 英文字抽出の有効性の評価

前節までに作成してきた英文字専用の切り出しと辞書とを組み合わせて英文字抽出の効果を実際の認識実験を行って検討する。ここで、有効性の比較対象として、日本語の切り出し (統合処理含む) と日本語の辞書を用いて、英文字抽出をしない場合 (以下、英文字抽出をしない認識と呼ぶ) を用意した。

4.4.1 実験方法

同じスキャナで読み込んだ文書を英文字抽出をした場合としない場合とでそれぞれ認識実験をする。ここで、切り出しと辞書以外の処理は同じ手法を用いた。即ち、切り出し後、正規化、細線化、線素化をほどこし、特徴量として方向線素特徴量を用いて、識別にはシティーブロック距離を用いた。

認識対象 LaTeX で 10pt、17pt の大きさで作成した英語が平均約 10%含まれている日本語文書 (計 1314 文字、そのうち英文字 154 文字) を用意した。

4.4.2 実験結果

各々の場合についての実験結果を以下の表にまとめた。

表 4.6: 英文字抽出の有効性の検討のための認識実験の結果

文字の大きさ	認識率			
	英文字抽出しない場合		英文字抽出した場合	
	全文章	英文字のみ	全文章	英文字のみ
10pt	94.82%	68.18%	92.85%	70.13%
17pt	94.67%	70.13%	93.38%	79.90%

まず、英文字だけの認識率に注目すると 10pt、17pt の両方で英文字抽出した場合の認識率が高くなっている。これは 英文字抽出の結果、英文字用の切り出しを用いて切り出しミスを少なくし、英語だけの辞書を利用して、識別候補を少なくすることによる認識率の向上と言える。よって、英文字認識だけを見る限りでは、英文字抽出は有効的であると言える。

次に、文章全体の認識率に注目すると10pt、17ptの両方で英文字抽出した場合の認識率が劣っている。本来、英文字の認識率が向上したので文章全体の認識率も向上するはずなのであるが、英文字抽出法自体がまだ不完全なものであるため、日本語を誤って英語とみなして抽出してしまった影響がここであらわれたと考えられる。そこで、それぞれの場合の日本語の誤抽出率を調べてみると、10pyの時に1.98%、17ptの時に2.50%であり、そもそもこれらの誤抽出文字が正しく抽出されていれば、英文字抽出が有効的となると予想される。

4.5 まとめ

認識実験の結果、今回提案した英文字抽出は不完全であったため、有効的であったとは言い難い。しかし、英文字の抽出をより高精度に実現できれば、英文字抽出法は有効的になるだろうと予想できる結果が英文字の認識率の向上から分かった。よって、より高精度な英文字抽出法の開発が今後の課題と言えよう。

また、本研究では、英文字の切り出しと辞書を簡易的なものを用いたが、これらも、英語OCR並みの性能が得られるものを作成し、英文字領域の認識に用いることができれば英文字抽出の効果も増していくだろうと思われる。これもこれからの課題の1つと言えよう。

第 5 章

結論

5.1 結論

本研究では、文字認識の前処理の段階の切り出しを正確に行うために、あらかじめ文書の中を英文字領域とその他の領域に分ける英文字抽出法を提案した。これにより、英語が混在する日本語文書に対しても、日本語の切り出しと英語の切り出しを各々に適した場所で使用できるため、切り出しの成功率を向上することができるようになった。

また、本研究で提案した英文字抽出法の性能を評価するために、簡易的な英文字用来り出しと辞書も作成した。英文字用切り出しにおいては、英文字切り出しで問題とされてきた斜体の切り出しにも対応できるようにするために射影分布を用いた斜め切り出しを取り入れた。また、英文字用辞書においては、フォントの数を 33 種類と大きくし、いろいろな書体に対応できるようにした。そして、英文字認識で問題とされている連字 (ligature) を標準パターンに組み入れることにより英文字の識別力を強化させた。

更に、本研究で提案した英文字抽出法の有効性を検証する認識実験も行った。しかし、この実験では、英文字だけに限った認識の時は有効であることが確認できたが、文章全体に対しての有効性は示すことができなかった。これは、英文字と誤って一部日本語も抽出してしまったことが原因であると分かった。この英文字抽出の際の誤抽出の原因としては、文字要素の英文字らしさの評価に用いるスコアを方形フィルタでフィルタリングするため、英文字領域の境界に日本語が残留してしまうことが分かった。

5.2 今後の課題

5.2.1 英文字抽出の高精度化

本研究で提案した英文字抽出法は、英文字の外観情報のみからあらかじめ文書の中の英文字部分を捜し出すものであった。そして、今回、英文字抽出に用いた英語の外観特徴は文字の横幅、ピッチのみであった。実験結果を見ても分かるように、このままではまだまだ英文字抽出は不完全である。よって、英文字の外観情報をもっと増やして (例えば、文字

の高さとその乱れ具合など) より高精度な抽出を実現することが必要である。

また、誤抽出の原因である英文字領域の境界での残留日本語については、スコアリングの際のフィルタリングを方形フィルタではなく、スコアの対象となる文字自身のスコアに重みを付けるような、言わば、ガウシアン的なフィルタを用いれば減少させることができるだろうと考えられる。

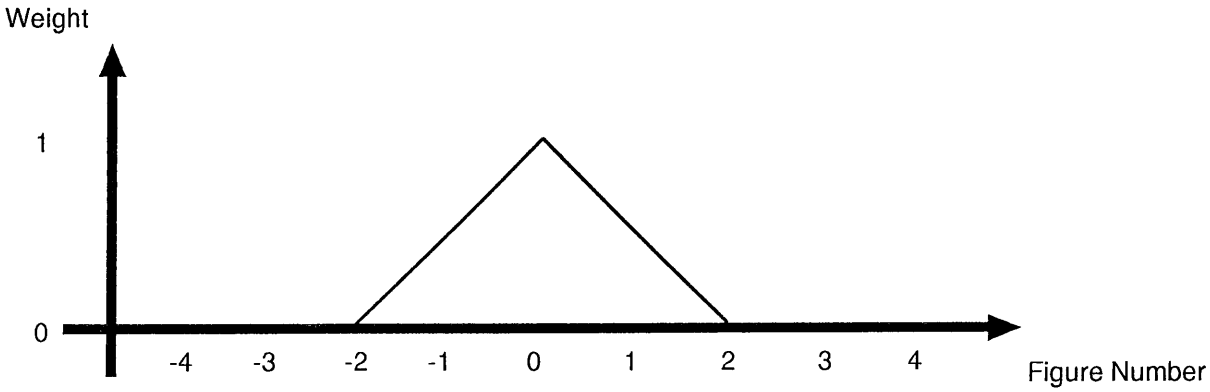


図 5.1: ガウシアン的フィルタの1例

更に、考えを一歩進めて、行単位で英文か日本文か英語が混在した日本文かの類別ができれば、英文字抽出を必要な時にしか用いなくてもよくなるため、より、英文字抽出法の効果が増すのではないかとも考えられる。

5.2.2 文字切り出し法の強化

本研究では、英文字抽出の性能の評価のために簡単な英語用の文字切り出し部を作成したが、この切り出し部を英語 OCR を参考にしながらもっと強力なものにできれば英文字抽出の効果を高め、文書全体の認識率も向上させることができると考えられる。

具体的な対策としては、射影分布による文字要素の分割を連結成分に着目した文字要素の分割に変更して、入り組み文字の分離を多くできるようにすることと、斜め切り出しも用いて、接触文字の切り出しを高精度に実現することが考えられる。

5.2.3 英文字用辞書の充実

本研究では、英文字用文字切り出しに加えて英文字用辞書も作成して英文字抽出の性能を検証するための実験を行った。しかし、辞書と言っても、集めてきたフォントを全て標準パターンとして羅列しただけのものに過ぎない。よって、クラスタリング等の処理を施して、辞書サイズをより小さくし、かつ、認識性能を向上させることが必要であると考えられる。

また、今回作成した辞書には連字 (ligature) も取り入れたが、これからも、新たな連字が発見される度に辞書に取り入れていかなければならないと考えられる。

謝辞

本研究を進めるにあたり、全般的に御指導頂いた東北大学工学部 阿曽弘具教授に心から感謝いたします。

そして、文字認識専門分野において多くの御指導、助言をしてくださった東北大学工学部阿曽研究室の後藤英昭氏、金津知俊氏、東北大学情報処理教育センターの大町真一郎助手に深く感謝いたします。

さらに、日頃から本研究につきまして御討議、御協力して頂きました阿曽研究室の皆様に深くお礼申し上げます。

参考文献

[1] 津雲淳, 浅井紘:" 文字認識技術の最近の動向", 電子情報通信学会技術研究報告, IE88-5, pp. 31-38(1988 年 4 月)

[2] 坂井邦夫:" 文字・文書の認識と理解", 電子情報通信学会誌, Vol. 71, No. 11, pp. 1182-1191(1988 年 11 月)

[3] 秋山照雄, 内藤誠一郎, 増田功:" 非接触文字優先切り出しによる印刷物からの文字切出し法", 電子通信学会論文誌 (D), Vol. J67-D, No. 10, pp. 1194-1201(1984 年 10 月)

[4] 馬場口登, 塚本正敏, 相原恒博:" 手書き日本文字列からの文字切り出しの基礎的考察", 電気通信学会論文誌 (D), Vol. J68-D, No. 12, pp. 2123-2131(1985 年 12 月)

[5] 南敏, 中村納:" 画像工学", テレビジョン学会編, (1989 年)

[6] 辻善丈, 浅井紘:" 分散最小基準に基づく適応型文字分離方式", 電子通信学会論文誌 (D), Vol. J68-D, No. 8, pp. 1497-1504(1985 年 8 月)

[7] 中川喜晶, 若林哲史, 鶴岡信治, 木村文隆, 三宅康二:" 英文 OCR における接触文字の切り出しアルゴリズム", 電子情報通信学会技術研究報告, PRU93-128, pp. 93-99(1994 年 1 月)

[8] 中村納, 鈴木弘之, 南敏:" 横書き日本語文書における個別文字の抽出", 電子通信学会論文誌 (D), Vol. J68-D, No. 11, pp. 1899-1909(1985 年 11 月)

[9] 佐藤道弘, 木田博巳:" 不定ピッチを含む印刷文書における文字切り出し手法", 電子情報通信学会技術研究報告, PRU88-159, pp. 97-104(1989 年 3 月)

[10] 有吉俊二:" 動的な仮説生成・検証による日本語印刷文書からの文字の切り出し", 電子情報通信学会技術研究報告, PRU93-47, pp. 33-40(1993 年 9 月)

[11] 大町真一郎:" 高速高精度文字認識システムに関する研究", 東北大学大学院工学研究科情報工学専攻 博士学位論文 (1993 年 3 月)

[12] 孫寧:" 文字認識の高速化と高精度化に関する研究", 東北大学大学院工学研究科情報工学専攻 博士学位論文 (1991 年 3 月)