1P-7

数式認識のための 高次情報を利用した文字誤認識訂正法の個別評価

瀧口祐介†, 岡田　稔†, 三宅康二††

† 早稲田大学大学院情報生産システム研究科, †† 中部大学工学部情報工学科

1　はじめに

科学技術系の書籍や論文に含まれる構成要素の一つとして，数式が挙げられる．数式を正しく認識・理解する技術が確立できれば，紙文書のデジタルアーカイブ化に有用である．さて，数式は，アルファベット，数学記号，ギリシャ文字および数字で構成され，それらの書体や字体と二次元平面上の配置で数式としての意味を表す．このため数式認識・理解を目的とする場合，文字認識精度の向上の重要性は高い．そのため筆者らは数式を対象とした認識システムの開発 [1] と，数式に関する高次情報を利用した誤り訂正法 (本訂正法)[2, 3] を提案している．

本稿では本訂正法による実験結果の個別評価によって訂正失敗の要因を考察し，その知見に基づいた今後の課題を述べる．なお本稿では，数式中に現れる文字・数学記号を単に文字，文字認識部で得られる正読文字の候補を候補文字と呼ぶ．

2　文字誤認識の訂正法

本訂正法ではまず，文字認識結果として得られた候補文字 n 個の相違度と数式の構造情報と数式中の隣接文字の接続方向を考慮した共起確率から，認識結果の数式らしさを表すコストツリーを生成する．その後，コストツリー中の全てのノードを一度だけ通るツリーのコストの総和を計算し，そのコストが最小となる組み合わせを認識結果として出力する．

2.1　数式キーワード

数式には関数名を始めとした様々なキーワードが存在する．そこで我々はこの様なキーワードを特に数式キーワードと呼び，これを利用して文字の誤認識の訂正を行う．具体的には，コストツリー内で水平方向に接続されたノードに注目し，それらに含まれる候補文字の組合せによってキーワード文字列の構成が可能かを調べ，可能であれば，そのキーワードを構成する全ての候補文字の相違度を減少させる．

2.2　接続方向別の共起確率

本訂正法ではコストツリーの重みの定義に，数式における隣接文字の，接続方向を考慮した共起確率を用いている．ここで共起確率とは，二文字が共に生起する確率である．また本研究では数式中の隣接文字の接続方向として，水平，右下，右上，下，上，左上，左下の 7 つを仮定している．

3　実験・考察

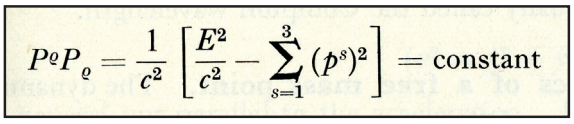
3.1　実験

約 50 年前に印刷された物理学の書籍 [4] から解像度 600dpi で読み込んだ 59 枚の数式画像，1,330 個の文字について実験を行った．訂正に用いる候補文字の数は $n = 5$ とし，共起確率は鈴木らによって提供されている数式画像データベース [5] から得たものに，経験的なバイアスを加えて用いた．ただし，本訂正法のみの評価を行うために認識順位 1 位の認識率が 87.4%の低精度な文字認識エンジンを用い，数式構造は正解を手入力した．実験の結果，本訂正法によって文字単位の認識率が約 87.4%から 92.4%に，数式単位の認識率が約 10.0% から 26.7% に改善したことを確認した．ただし文字単位および数式単位の認識率とはそれぞれ，総文字数に対して字種と字体が正しく認識された文字 (正読) の割合と，総数式数に対して数式に含まれる全ての文字が正読である割合である．図 1 は本訂正法によって全ての誤読文字が正しく訂正できた例である．図 1(a) は入力画像を，(b) と (c) はそれぞれ，訂正なしとありの結果を LaTeX で清書したもののプレビューを表している．一方，図 2 は誤読文字が正しく訂正できなかった例である．図 2(b) では，‘2’ が ‘z’ に (**失敗 A**)，‘σ’ が C に (**失敗 B**)，分数線が ‘$\rightarrow$’ に (**失敗 C**)，それぞれ誤って認識されており，(c) でも正

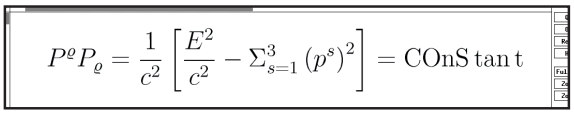
Individual Evaluation of Character Recognition Error Correction Method for Mathematical Formulae Recognition using Higher Level Information

†Yusuke TAKIGUCHI and Minoru OKADA, Waseda University

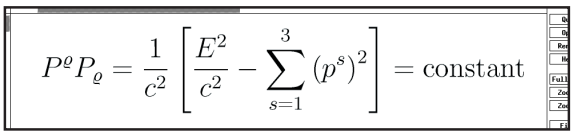
††Yasuji MIYAKE, Chubu University



(a) 入力画像 (1,481 × 290 [pixel])

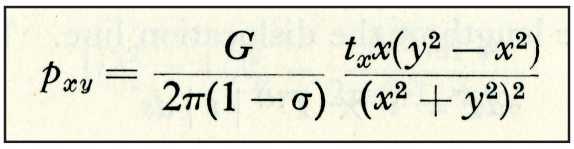


(b) 認識結果 (誤り訂正なし)

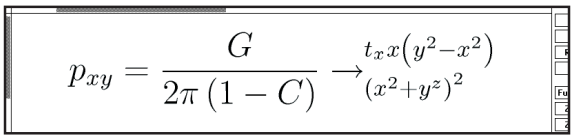


(c) 認識結果 (誤り訂正あり)

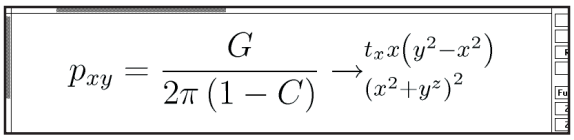
図 1: 実験結果 1 (訂正成功)



(a) 入力画像 (1,134 × 271 [pixel])



(b) 認識結果 (誤り訂正なし)



(c) 認識結果 (誤り訂正あり)

図 2: 実験結果 2 (訂正失敗)

しく訂正できていないことが確認できる．

3.2 考察

失敗 A の原因は，誤認識を生じた '2' における文字の認識結果の上位 5 位以内に，正読文字が存在していなかった (実際には第 7 位) ためである (**原因 A**)．また**失敗 B** の例では，誤認識を生じた 'σ' と '$-$' の共起確率は 'C' と '$-$' のものより高かったが，文字認識処理で出力された 'σ' の相違度が大きかったため，正しく訂正できなかった (**原因 B**)．一方，**失敗 C** の場合，誤認識を生じた分数線とその隣接文字との共起確率に差が無かったために，誤認識が訂正されなかった (**原因 C**)．上掲した失敗の**原因 A**，**B**，**C** から考えられる失敗要因は次の通りである．

- **要因 α**: 正読文字が候補文字の中に存在しない
- **要因 β**: 正読文字のコストが高い
- **要因 γ**: 共起確率の学習サンプルの不足・偏り

この内，**要因 α** については，文字認識処理の精度向上によって正しく認識，訂正成功もしくは**要因 β**，**γ** による訂正失敗に結果が遷移することが予想される．本実験では低精度な文字認識エンジンを用いたためにこのような失敗が起こったが，これは文字認識精度のみの問題であり，本訂正法固有の問題ではない．一方，**要因 β** と **γ** による失敗について考えた場合，より多くの学習サンプルから有意な共起確率を得る必要がある．ただし字種単位で共起確率を得ると大量のサンプルが必要となり現実的では無い．そのため数学的意味で字種を分類することで，必要な学習サンプル数を減少させる必要がある．また，本実験の結果において，**失敗 C** のように分数線を誤認識したものが幾つか見られた．この原因は文字認識処理の際に認識対象文字のサイズを正規化しているためである．これについては，数式内の隣接文字において，文字の縦横のサイズ比を考慮して認識を行うことが有効であると考えられる．

4 まとめ

本稿では，数式を対象とする文字誤認識の訂正法の個別評価の結果を示した．特に訂正失敗の具体例を示し，失敗の要因ついて考察した．考察より，失敗要因への対応策として，数学的意味を考慮した字種の分類，隣接文字同士のサイズ比の考慮，の必要性を確認した．今後は上掲した課題に取り組む．

参考文献

[1] Y. Takiguchi, M. Okada, and Y. Miyake: "A Fundamental Study of Output Translation from Layout Recognition and Semantic Understanding System for Mathematical Formulae", *Proc. of ICDAR2005*, pp. 745–749 (2005)

[2] 瀧口祐介, 岡田　稔, 三宅康二: "高次情報を利用した数式文字認識の誤り訂正法の一検討", 信学技報, PRMU2005-248, pp. 107–112, 6 pages (2006)

[3] 瀧口祐介, 岡田稔, 三宅康二: "共起確率行列を用いた数式文字認識の誤り訂正法の評価", 情報科学技術レターズ (*IT Letters*), 4F-5, 4 pages (2006)

[4] D. H. Menzel: "Fundamental Formulas of Physics", Prentice-Hall, Inc. (1955)

[5] M. Suzuki, S. Uchida and A. Nomura: "A Ground-Truthed Mathematical Character and Symbol Image Database", *Proc. of ICDAR 2005*, pp. 675–679 (2005)