修士学位論文要約（令和４年３月）

日本語・中国語文字の高速・高精度な認識に関する研究

松末　三星
指導教員：菅沼　拓夫，　学位論文指導教員：後藤 英昭

Fast and Accurate Japanese/Chinese Character Recognition

Mihoshi Matsusue
Supervisor: Takuo SUGANUMA,　Research Advisor: Hideaki GOTO

Demand for character recognition has increased in recent years. In the background, there is a widespread use of smart phones and other devices, and a growing shortage of labor. However, character recognition includes some difficult tasks. One of them is the character recognition with a large number of variants which is commonly found in Japanese and Chinese. Recognition of these characters requires a lot of computation. To address the task, Hierarchical Overlapping Clustering (HOC) with Linear Discriminant Analysis (LDA) was proposed. The method is a for candidate reduction using a binary tree dictionary. In my study, a new candidate reduction method based on the HOC is developed by combining some leaf nodes that have been traced multiple times under different conditions in order to achieve more efficient candidate reduction. Experimental results show that the coverage is improved from the previous 97.4% to 97.6% and speedup ratio is improved from the previous 6.8 to 7.3.

1. はじめに

現代では，パソコンやスマートフォン，タブレット端末の普及が大きく進み，それらの電子機器上で文字情報を扱うアプリケーションの需要が大きく増えてきている．また，世界的にも少子化が進んでいることから，書類の処理などに必要となる手間を削減することのできる文字認識技術は今後も重要性を増すと考えられる．

文字認識では，最近傍探索を用いた全数整合が広く用いられている．文字認識はまず，予め用意された学習データを用いて辞書データ（テンプレート）を作成する．次に，クエリから抽出された特徴ベクトルとテンプレートを比較することで認識を行う．最近傍探索は，日本語や中国語などの字種数が多いものに適用する場合，計算量が大きくなる問題がある．本研究では，日本語・中国語の各データセットを対象に，階層型重複クラスタリングにより構築された二分木型辞書 [1]を異なる条件下で複数回走査することで，速度を維持しつつ高精度化する手法について検討した．

2. 先行研究

先行研究 [1]である階層型重複クラスタリング(HOC, Hierarchical Overlapping Clustering)は，線形判別分析(LDA, Linear Discriminant Analysis) [2]を用いることで二分木型辞書を構築し，それを用いることで文字認識の候補を削減する手法である．辞書作成段階では，クラスタcに含まれる学習データに LDA を適用し，最も大きい固有値に対応する固有ベクトル$\boldsymbol{v}_c$を射影軸として，式(1)に従い，終了条件式(2)のいずれかを満たすまでクラスタの分割を再帰的に行う．射影値の平均をH_c，次層の 2 つのクラスタに含まれる字種数をそれぞれ$N_{c,1}, N_{c,2}$とする．また，外れ値の存在によって分割が妨げられるのを防ぐため，射影値がクラス内平均から遠い順に$\lfloor \beta \cdot S_i \rfloor$個のサンプル数を拾い外れ値として，それらを外れクラス$O_{i,c}$として次層に登録する．

$$\begin{cases} \max\limits_{j \in C_i - O_{i,c}} r_{i,j,c} & < & H_c - \alpha \cdot \sigma_i \\ \min\limits_{j \in C_i - O_{i,c}} r_{i,j,c} & > & H_c + \alpha \cdot \sigma_i \end{cases} \tag{1}$$

K_1はリーフノード内のクラス数，C_Rは子ノードとの重複率に関するパラメータである．

$$\begin{cases} K_1 & \geq & N_c \\ C_R & \leq & \max\left(N_{c,1}, N_{c,2}\right) / \mathrm{N_c} \\ H_T & \leq & (\text{depth of the cluster } c) \end{cases} \tag{2}$$

3. 階層型重複クラスタリングノード併合法

階層型重複クラスタリングによって構築された二分木辞書を，複数の異なる条件下において複数回走査することによって高精度化する手法を検討する．従来手法では，候補削減の段階でカバー率（走査して到達したリーフノードに正解字種クラスが含まれている割合）が 95~97%と，詳細分類の前の段階で正解字種の多くを除いていた．そこで，本研究ではツリーの走査を行う際に条件を追加することによって通常ではたどり着けない，正解字種が含まれているノードへの到達を図る．

この手法ではまず通常通り各ノードで分割点とクエリの大小比較を行うことでツリーの走査を行う．リーフ

ノードまでたどり着いたら，もう一度走査を最初からやりなおす．この際に条件を追加することで一度目の走査とは異なるリーフノードへの到達を図る．たどり着いた複数のリーフノードを結合して新たなリーフノードを作成する．処理の概略を図 1 に示す．HOC で作成された二分木辞書を複数回走査する手法を以下，階層型重複クラスタリングノード併合法とする．本稿では追加するルートの変更の条件として，分割点とクエリの距離が最も大きいときにルートを変更する max，最も小さいときに変更する min，平均値より小さいときに変更する ave，ルートノードで変更する first，リーフノードの一つ上で変更する last の五つを比較する．

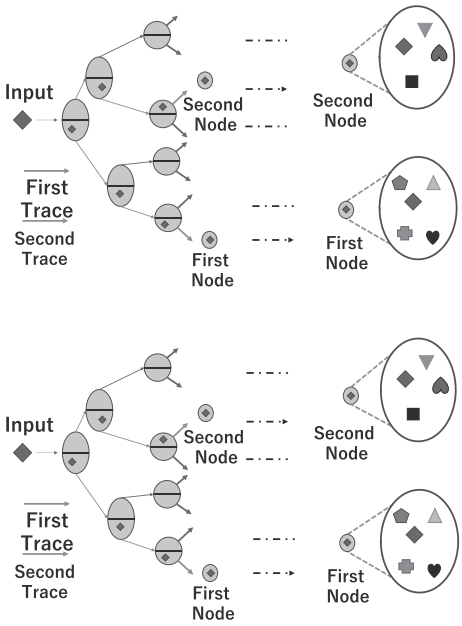

図 1 階層型重複クラスタリングノード併合法概要

4. 評価実験と考察

開発した手法を評価する実験を行った．従来手法との速度・カバー率の比較を行った．対象とするデータセットは中国語のデータセットである HCL2000 [3] を採用した．また，考慮する必要のあるパラメータとして，リーフノードのパラメータであるK_1がある．新しい手法ではリーフノードを二つ結合させるため，従来手法と探索する字種数をできるだけ揃えられるよう，従来手法は$K_1 = 600$，新手法は$K_1 = 300$を採用した．図 2 にカバー率の比較のグラフを，図 3 に速度向上比の比較のグラフを示す．従来手法は Odate と表記する．カバー率，速度向上比の双方のグラフにおいて，ave と min が従来手法を上回る結果が得られ，新手法はカバー率と速度の面において従来手法を改善したことが確認できた．ave と min においてカバー率が向上した理由としては，分割点とクエリとの距離が近いときにルートの変更を行ったことで，従来手法ではたどり着くことができなかったノードが候補に含まれるようになったためと考えられる．

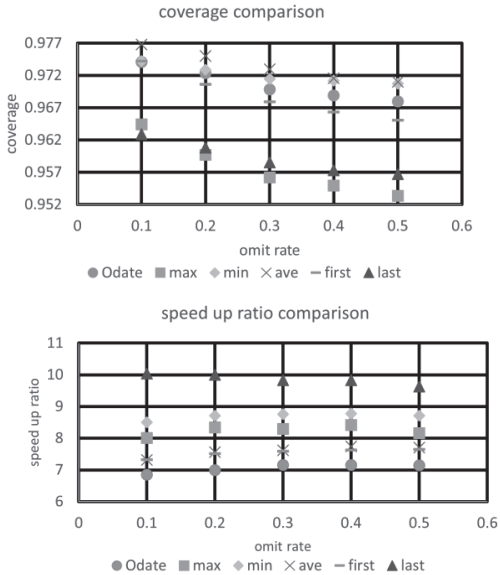

図 3　速度向上比の比較(HCL2000)

5. まとめ

本研究では，日本語・中国語の手書き文字認識の速度を維持した高速化のために，二分木辞書を異なる条件下で複数回走査する階層型重複クラスタリングノード併合法を開発した．また，評価実験を行い，新手法が速度を維持しつつカバー率を従来手法の 97.4%から 97.6%に改善したことを確認した．

文献

1)R. Odate and H. Goto, “Highly-accurate fast candidate reduction method for Japanese/Chinese character recognition,” in Proc. ICIP 2016, pp.2886-2890,Phoenix, AZ, USA, Sep. 2016.

2) P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, “Eigenfaces vs Fisherfaces:Recognition using Class Specific Linear Projection, ”IEEE Transactionson Pattern Analysis and Machine Intelligence, vol.19, no.7, pp.711-720, Jul. 1997.

3) H. Zhang and G. Chen and C. Li, “HCL2000 - A Large-scale Handwritten Chinese Character Database for Handwritten Character Recognition,” Proc. ICDAR 2009, pp. 286–290, 2009.