2K-07

# 甲骨文字データベースの構築と候補テンプレートの検索

柴田睦月† 辻翼† 紙徳直生† 山形卓也† 孟林† 山崎勝弘†
立命館大学 理工学部†

## 1. はじめに

甲骨文字認識は、文字の起源や歴史の解明、古代文献の解読に重要であるが、劣化などが原因で認識が難しい。本研究では、現在までに確認されている甲骨文字のデータベースを構築し、そこから候補テンプレートを検索して、認識することを目指す。本論文では、甲骨文字の特徴量に注目し、特徴点(端点、分岐点など)、特徴点間を繋ぐ線の本数、文字の 9 区分ごとの面積分布を調査し、甲骨文字データベースを構築する。そして、甲骨文字の自動認識を行う際に文字の特徴量を用いてデータベースを検索し、認識対象の甲骨文字原画像と類似している候補テンプレートを抽出してマッチングを行う。

## 2. 甲骨文字認識システム

甲骨文字認識システムを図 1 に示す。本システムは、ノイズ除去処理、特徴抽出、文字認識により構成される。

まず、認識対象である甲骨文字原画像[1]にガウシアンフィルタ、2 値化処理を行い細かなノイズを除去し、ラベリング処理で一定の大きさのノイズを除去する。次に、細線化処理を行った後、ハフ変換で直線を抽出する。その後正規化処理を行い文字の傾きや大きさなどを修正し、骨格を抽出する。正規化された原画像から特徴点、線の本数、面積分布といった特徴量を取得してデータベース内のテンプレートの特徴量と比較し、類似度が高いテンプレートを抽出する。抽出されたテンプレート画像とのマッチング処理で類似度を算出し、文字の認識を行う。

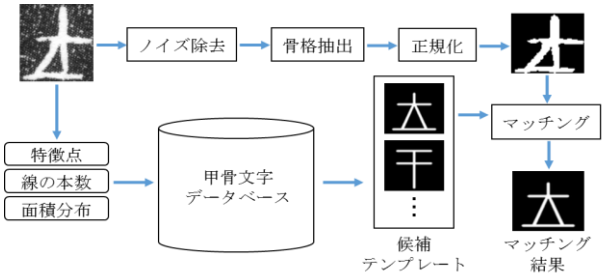



図 1　甲骨文字認識システム

## 3. 甲骨文字データベースの構築

現在甲骨文字データベース[2]には現代漢字とテンプレートが登録されているが、画像処理での甲骨文字認識に直接使用できない。本研究ではこれを元に甲骨文字認識用のデータベースを構築し、画像処理を用いた自動認識を目指す。

Construction of Oracle Bone Inscriptions Database and Retrieval of Candidate Templates
Mutsuki Shibata†, Tsubasa Tsuji†, Naoki Kamitoku†, Takuya Yamagata†, Lin Meng†,and Katsuhiro Yamazaki†
†College of Science and Engineering, Ritsumeikan University.

### 3.1 甲骨文字データベースの構成

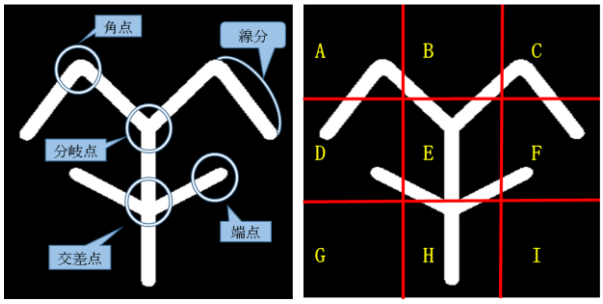
図 2 に甲骨文字データベースの構成を示す。甲骨文字 1 文字のレコードには、現代漢字、テンプレート画像[2]、及び特徴量が含まれており、現在約 1000 種類の甲骨文字を登録している。特徴量には各特徴点数、線の数、及び 9 分割した面積の割合が含まれる。また、各文字に対する原画像ファイルを別に用意している。

| 番号 | 漢字 | 画像 | 端点 | 角点 | 分岐点 | 交差点 | 総数 | 線分 | 面積 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | A | … | I |
| 0 | 貞 |  | 2 | 2 | 10 | 0 | 14 | 18 | 13.7 | … | 18.3 |
| 1 | 癸 |  | 8 | 0 | 4 | 1 | 13 | 12 | 5.5 | … | 13.2 |
| 2 | 王 |  | 5 | 0 | 2 | 1 | 8 | 8 | 0 | … | 12.1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

図 2　甲骨文字データベースの構成

### 3.2 甲骨文字の特徴量

甲骨文字の特徴量は図 3 に示すように特徴点の数、線の本数、面積分布の 3 つに分類される。




特徴点数[個]
端点　：5　角点　：2
分岐点：1　交差点：1
線の本数[本]：8

面積分布[%]
A：11.2　B：8.8　C：11.2
D：11.5　E：26.6　F：11.5
G：0　H：16　I：0

図 3　甲骨文字の特徴量

特徴点は端点、角点、分岐点、交差点に分類される。端点は筆画の終端(あるいは始点)、角点は筆画が折れた点、分岐点は筆画同士が繋がっており、かつ交差していない点、交差点は 2 つ以上の筆画が交差している点と定義する。各文字に対してこれらの点の個数を登録する。

線の本数は文字の画数ではなく、文字の特徴点間を繋ぐ線の本数を登録する。また、甲骨文字画像を 9 つの区分(A~I)に分割し、各区分に存在する文字の面積の割合を算出し、1 文字に対してこれら 9 つの区分ごとの面積割合を登録する。これらの特徴量を検索キーとして、原画像に類似した候補テンプレートを検索する。

## 4. 候補テンプレートの検索

### 4.1 特徴点と線による検索

原画像から取得した端点、角点、分岐点、交差点やそれらの特徴点間を繋ぐ線の本数をデータベース内の全てのテンプレートの特徴量とそれぞれ比較する。文字の歪みや劣化などで原画像からの特徴点の抽出が完全ではないことを見込み、各要素の誤差を考慮して検索を行い、類似する特徴量を持つテンプレートをマッチング候補テンプレートとして抽出する。

### 4.2 面積分布による検索

4.1 の手法で抽出された候補テンプレートの中には文字の形状が大きく異なるものも含まれている。この面積分布による検索では、文字の形状が原画像に対して大きく異なるテンプレートを候補から省くことを目的とする。

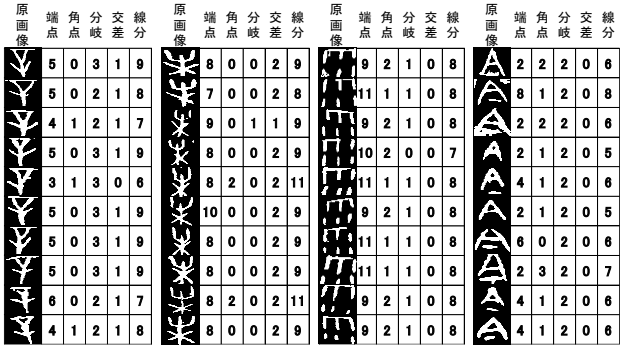
## 5. 実験と考察

### 5.1 実験内容

甲骨文字原画像の特徴点数を用いてデータベース検索を行い、検索抽出数と正答テンプレートの有無を調査する。また、その中から線分の数が一致しているテンプレートを抽出し、さらなる絞り込みを行う。4 文字 10 画像計 40 枚の原画像の特徴点を取得し、候補テンプレートを検索する。

### 5.2 結果

40 枚の原画像の特徴点数と線分数を図 4 に示す。




“辛” “未” “雨” “今”

図 4　甲骨文字原画像の特徴量

これらの原画像の特徴点数を用いた検索結果について、特徴点数の誤差±0、±1 の場合は、正解テンプレートが絞られたのが半分以下であり、誤差±2 の場合は、全ての “辛”,”未”,”雨”では正解テンプレートが絞られ、それぞれの候補テンプレートの平均文字数は”辛”206 文字、”未”139 文字、”雨”96 文字となった。”今”では、2 番目の文字除いて正解テンプレートが 189 文字に絞られた。

また、特徴点数による検索で絞られたテンプレート画像から、線分が一致する文字を絞り込む。”雨”の特徴点数と線分の数による検索結果の一例を図 5 に示す。特徴点数のみの検索では、それぞれ 134 文字、51 文字となっているが、そこから線分の数が一致するものを抽出すると、それぞれ 15 文字、6 文字と大幅に候補テンプレートを絞り込むことができた。また、双方の検索結果に”雨”の正解テンプレートも含まれている。

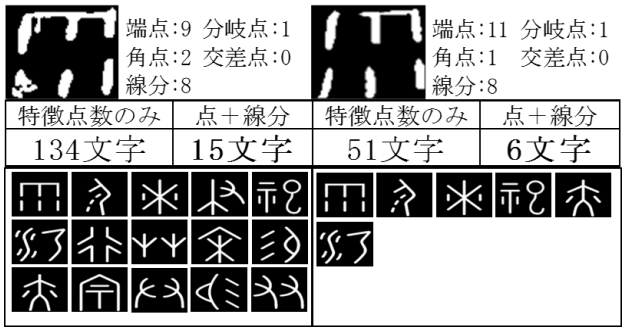



図 5　特徴点数と線分の数による検索結果

### 5.3 考察

“今”の 2 番目の画像のように大きなノイズが残り、特徴点数が大幅に異なる画像を除くと、誤差±2 での検索で約 1000 文字の中から正解テンプレートを含む候補テンプレートを約 100~200 文字に絞り込むことができた。また、”雨”での実験では、線分の数が一致しているテンプレートを抽出すると約 10 文字と、データベース内の約 1%まで候補を絞り込むことができた。候補テンプレートを少なく、かつ正確に絞り込むために、以下のような改良の余地がある。

(1) 検索条件の優先度の変更

本実験では特徴点数、線分の本数の順に検索を行ったが、線分の本数、特徴点数の順で検索を行った検索結果や、他の特徴量と併せて行った検索結果を調査し、重要度の高い検索条件を見定める必要がある。

(2) 線分の差±1 の文字の抽出

認識対象の原画像にわずかなノイズやひげ、欠損などが生じると、特徴点と同様に線分の本数も差異が生まれる。そういった問題に対応するために線分の差±1 までのテンプレートを抽出することで正解テンプレートが検索結果から外れる可能性を削減できると考えられる。

(3) 面積分布などの特徴量を併せた検索

特徴点数や線分の数を用いた検索に 4.2 の面積分布を追加して検索を行い、形状の大きく異なる文字を検索結果から除くことで抽出文字数をさらに絞り込むことができると考えられる。

## 6. おわりに

本研究では甲骨文字 1 文字ごとのテンプレート画像を用意し、文字の特徴量から甲骨文字データベースを構築した。また、原画像から取得した特徴量を用いて類似するテンプレートをデータベース内から検索し、候補テンプレートとして絞り込むことができた。今後の課題として、原画像からの特徴量自動抽出の精度向上、候補テンプレートの検索条件の最適化などがあげられる。

参考文献

[1]濮茅左, 上海博物館蔵甲骨文字, 上海辞典出版社, 2009.
[2]落合淳思, 甲骨文字データベース,
http://koukotsu.sakura.ne.jp/top.html