1N-08

特徴量を用いた甲骨文字の候補テンプレート抽出と認識

石井　康史†　　藤川　佳之†　　孟　林‡　　山崎　勝弘‡

立命館大学大学院理工学研究科†　　立命館大学理工学部‡

1. はじめに

甲骨文字は、3,000年以上前の中国殷代に使われた亀の甲羅や獣の骨に刻まれた象形文字であり、漢字の祖形とも言われている。甲骨文字の解読は、我々の使う漢字の起源を知る上で非常に重要であるが、長い年月による劣化などが原因で認識しにくい問題がある。また、現在の解読は、歴史学者の経験や文脈と、従来からある文献によって行われることが多い。

従来の認識では、原画像の文字に対して、正答テンプレートを用意して認識を行い、認識精度の向上を図ってきた[1]-[3]が、文字を認識するためには、大量のテンプレートを含む甲骨文字データベースから正答テンプレートを自動抽出する必要がある。そこで、あらかじめ認識対象とデータベース内の全テンプレートに何らかの特徴量を付与し、その特徴量を用いて、データベースから認識の候補となる小数の候補テンプレートを抽出することで、認識の高速化と精度向上を図る。

2. 甲骨文字の特徴量

2.1 認識の流れ

従来の甲骨文字認識では、甲骨の拓本から切り出した認識対象となる原画像に対して、ガウシアンフィルタ・2値化・ラベリングによるノイズ除去、及び細線化とハフ変換による骨格抽出を行い、その骨格を正答テンプレートとマッチングすることで類似度を算出し、認識の判断を行ってきた[1][2]。

今後の認識では、甲骨文字データベースの中から候補となるテンプレートを自動抽出し、それらの類似度を同時に計算して認識を行う。

2.2 特徴量の定義

甲骨文字データベースから候補テンプレートを抽出するためには、認識対象とテンプレートに何らかの特徴量を付与する必要がある。

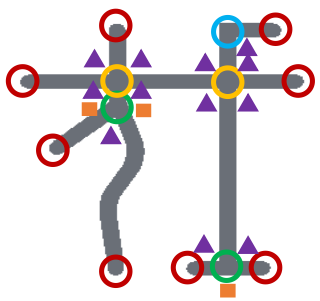
本研究では、その特徴量として特徴点数、線の本数、線の角度の数を定義する。特徴点は、文字に含まれる線の端点、角点、分岐点、交差点を抽出した総数である。線の本数は、抽出した特徴点を用いて、特徴点間に存在する線を 1 本と数えた総数である。また、線は点から点までの区間であるため、直線も曲線も含まれる。線の角度は、角点、分岐点、交差点の周辺に存在する角度を鋭角か直角（0～90 度）、鈍角（91～180 度）の 2 種類で抽出し、それぞれの総数を特徴量とする。

Extraction and Recognition of Candidate Template in Oracle Bone Inscriptions Using Feature Value

Koji Ishii †, Yoshiyuki Fujikawa †, Lin Meng ‡, Katsuhiro Yamazaki ‡

† Graduate School of Science and Engineering, Ritsumeikan University.

‡ College of Science and Engineering, Ritsumeikan University.



種類	個数
端点	8
角点	1
分岐点	2
交差点	2
特徴点	13
線	12
0～90度	12
91～180度	3

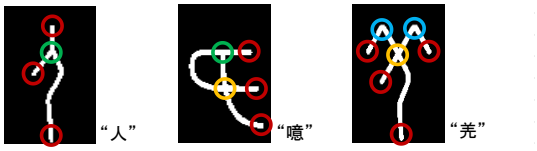
図 1　甲骨文字の特徴量

図 1 は甲骨文字の特徴量を抽出した例である。

3. 候補テンプレート抽出

3.1 甲骨文字データベース

本研究では、認識に用いるテンプレートを約2,000種類用意している。テンプレートは、本学文学部で開発された、文字の傾きがなく、線分の太さと文字の大きさが統一された甲骨文字画像である。これらのテンプレートから特徴量を抽出し、甲骨文字データベースを作成している[3]。甲骨文字データベースの構造を図 2 に示す。データベースに保存されているこれらの特徴量を認識対象より抽出した同様の特徴量を比較することで候補テンプレートの抽出に応用する。また現在、甲骨文字データデースには約 100 文字の特徴量を保存している。



文字	○	○	○	○	特徴点	線	0～90度	91～180度
人	3	0	1	0	4	3	1	2
噫	3	0	1	1	5	5	4	3
羌	4	2	0	1	7	6	4	2

図 2　甲骨文字データベースの構造

3.2 特徴量を用いた抽出

特徴量を用いた候補テンプレートの抽出と認識の流れを図 3 に示す。原画像からハフ変換までの処理によって、骨格を抽出した画像から、HOG における輪郭角度計算を用いて特徴点を抽出する[4]。抽出した特徴点から、線の本数、角度を抽出し、それぞれを文字の特徴量とする。さらに、抽出した特徴量を用いて、甲骨文字データベースに存在する全テンプレートの特徴量とのマッチングを行い、認識対象の特徴量と類似する特徴量を持つ候補テンプレートを自動的に抽出する。

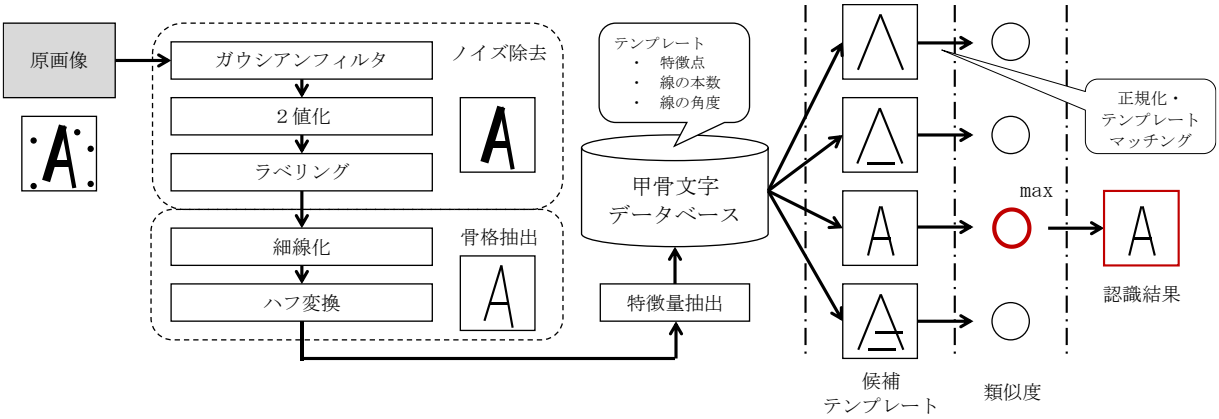


図 3　特徴量を用いた候補テンプレートの自動抽出と認識

3.3　候補テンプレートを用いた認識

抽出した全候補テンプレートと認識対象で、それぞれ正規化[2]を行い、テンプレートマッチングにより比較し、類似度を算出する。最終的に、類似度が最も高くなったテンプレートを認識結果として出力する。候補テンプレートを用いた認識により認識の高速化と高精度化を図る。

4. 実験

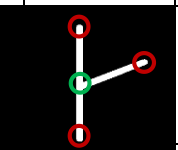
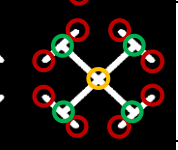
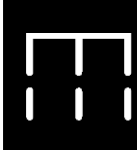
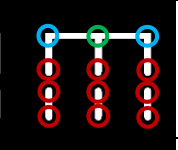
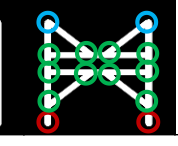
4.1　実験内容

特徴点を抽出した 15 文字のテンプレートに対して、特徴量である線の本数と角度の抽出を行い、その正確性を確認する。

4.2　実験結果

特徴量を抽出した例を表 1 に示す。テンプレート画像から特徴点を抽出した画像を用いて、テンプレートの線の本数と角度を抽出している。実験から、15 文字全てのテンプレートにおいて、特徴量が目視で確認できる特徴量と一致していることから、特徴量が正確に抽出できていることが分かる。

表 1　特徴量抽出の結果

テンプレート画像	特徴点抽出	特徴点数	線の本数	線の角度	
				0～90度	91～180度
		4	3	1	2
		13	12	12	4
		12	8	4	1
		14	18	16	16

4.3　考察

今回の実験で用いたテンプレートは文字の傾きがなく、線の太さや大きさが統一された画像であるため、特徴量の抽出が比較的容易であり、特徴量をうまく抽出することができた。候補テンプレート抽出では認識対象となる原画像の特徴量も抽出する必要がある。また、原画像とテンプレートの特徴量には誤差が生じるため、その誤差をどのようにして考慮するかを今後の実験によって誤差の統計をとり検討する必要がある。

特徴量を用いて候補テンプレートを抽出して認識を行うには、候補テンプレートの中に正答テンプレートが含まれているか、正答テンプレートの類似度が最も高くなるかが重要となるため、今後の実験によって検証する。

甲骨文字を候補テンプレートを用いて認識するためには、約 2,000 文字全ての特徴量をデータベースに保存することが必要となる。

さらに、甲骨文字認識は CPU で行っているが、今後は、GPU に認識の各処理を並列に動作するように実装し、認識をより高速化することも検討している。

5. おわりに

本研究では、甲骨文字認識の高速化と高精度化のために、特徴量を用いた候補テンプレートの抽出を提案した。抽出に用いる特徴量として、特徴点数、線の本数、線の角度を定義した。実験では、15 文字のテンプレートに対して特徴量の抽出を行い、全ての画像において正確に抽出できることを確認した。

今後の課題としては、特徴量を用いた候補テンプレートの自動抽出、甲骨文字データベースの拡張、GPU を用いた複数文字の同時認識などが挙げられる。

参考文献

[1] 孟, 石井, 藤川, 落合, 泉, 山崎, “甲骨文字認識プロジェクト”, 電子情報通信学会総合大会, D-12-11, 2015.
[2] 石井, 藤川, 孟, 山崎, “アフィン変換による正規化を用いた甲骨文字の認識率向上”, 情報処理学会関西支部支部大会, G-02, 2015.
[3] 石井, 藤川, 孟, 山崎, “甲骨文字認識における文字データベースの作成”, 電子情報通信学会総合大会, D-12-10, 2015.
[4] 藤川, 石井, 孟, 山崎, “HOG における輪郭角度計算を用いた甲骨文字の特徴点抽出”, 情報処理学会第 78 回全国大会, 1N-07, 2015.