CNNと文字のアスペクト比を用いたくずし文字認識

上田 佳祐[1]，薗頭 元春[2]，飯山 将晃[3]*

[1] 洛南高等学校，[2] 京都大学大学院情報学研究科，[3] 京都大学学術情報メディアセンター

要旨

紙媒体や画像上の文字を計算機により自動認識する文字認識の技術は，ビジネス文書などの自動処理だけでなく人文科学の研究などにおいてもデータ処理にかかる人的資源を削減できるという点で有用である．本研究では古典文献などのくずし文字の認識を目的とし，現代文字の認識で高い精度が得られている畳み込みニューラルネット（CNN）による識別と，くずし文字特有のアスペクト比を用いた識別を組み合わせた認識手法を提案する．実験の結果，提案手法により，単にCNNを用いた場合より高い精度の認識性能が得られた．

重要語句：文字認識，畳み込みニューラルネット，アスペクト比

1. 序論

過去の日本文化の研究においては過去の文字が読めることが必要であるが，明治から大正にかけて行われた日本語の標準語化の影響もあり，現代の日本語と江戸以前の日本語には大きな開きがある．そのため今日，日本の古典文学を専門家以外が読むことはできず，また，これらの文学作品が電子化されずに埋もれてしまっているという現状がある．この電子化の障壁となっているものの一つが，古典文学が現代人に容易に理解できないくずし字で書かれていることである．

この問題に対し，計算機によるくずし文字認識が解決のアプローチとして考えられる．しかしながら，現在の文字認識技術は現代文に対応したものがほとんどであり，くずし字に対する精度は現代のひらがななどと比べて高くない．その理由として，古典文学におけるくずし文字が現代文字ほど規格化されておらず，同じ文字でも見た目が大きく異なり，他の文字とのはっきりとした区別が困難であることが考えられる．

今日，文字認識分野においてはCNN（畳み込みニューラルネットワーク）[1, 2]を用いたアルゴリズムが非常に高い精度を達成している．しかしこのアルゴリズムは入力として固定サイズの画像しか扱うことが出来ず，そのサイズ比と合わない画像は縦や横に引き伸ばされてしまい，文字認識に対して精度が落ちてしまう．この傾向は，一つ一つの文字のアスペクト比が大きく異なるくずし字には顕著に表れてしまう．

そこで，文字のアスペクト比に着目した別の識別器をCNNと併用することによりこの問題を解決することを目指す．

2. くずし字認識問題

本研究では，ラベル付きの教師データとして人文学オープンデータ共同利用センターが公開している日本古典籍字形データセットを利用する．それらは，くずし文字一つ一つを切り出した画像とその文字のユニコードがペアとなっているおり，文字はすべてひらがな（46種）である．図1にくずし文字の例を示す．図1中で赤色で囲まれた範囲の画像と，それに対応するユニコードが与えられる．本研究では，この教師データ約5万文字分を用い，入力された未知の文字画像から，その画像に対応するひらがなの認識を出力する文字認識の構築を行う．

3. 提案手法

3.1 CNNを用いた文字認識

CNNは脳のニューロンの仕組みから着想を得たニューラルネットワーク（NN）と呼ばれる機械学習手法の一種である．提案手法ではCNNを用いて文字認識を行う．古典的なNNでは3層のノードからなり，層間のノード同士の枝の重みを学習する全結合層のみで構成されている

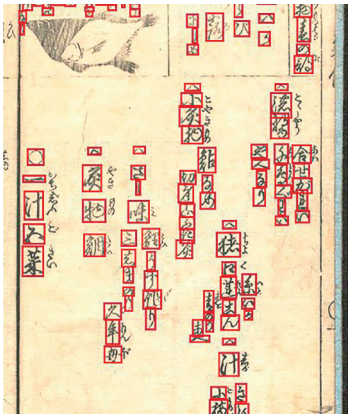

図1．くずし文字の例[3]

*内容に関する連絡先：iiyama@mm.media.kyoto-u.ac.jp

が，CNNではこれらに加えて畳み込み層やプーリング層などの画像の幾何構造に対応する特徴を抽出するための層も用いてネットワークが構築されている．

本研究で用いたCNNの構造を図2に示す．入力はRGB3チャンネルの固定サイズ（64 × 64）画像とし，学習時・認識時ともに，元の入力画像をこのサイズとなるよう拡大・縮小したものをCNNに与える．

畳み込み層では入力としてマルチチャンネルの2次元画像を受け取り，それに対してフィルタを畳み込む処理を行う．入力がCチャンネル，W × Hの画像とし，フィルタをCチャンネルの(2w+1)×(2h+1)サイズとした場合，その出力は1チャンネルの2次元画像となり，その値は次式で与えられる．

$$I'(s,t) = \sum_{c}^{C} \sum_{i=-w}^{w} \sum_{j=-h}^{h} F(c,i,j) I(c, s-i, t-j)$$

ここで，$I'(s,t)$ は出力，$F(c,i,j)$ はフィルタ $I(c,s,t)$ は入力である．この出力に対し，さらに活性化関数relu() を適用し，次の層への入力とする．なお，reluとはReLU（Rectified Linear Unit）のことであり，入力をxとしてmax(0, x)で表される．

また，この出力はフィルタで抽出される特徴が画像内のどこに表れているかを示しており，値が1に近いほど一致の度合いが強く，–1に近いほどネガに対する一致が強いことを示している．つまり，畳み込み層では画像内からフィルタに合う特徴を抽出していると見なすことができる．

プーリング層においては，先ほどの畳み込み処理後の値を入力として受け取り，ある一定サイズのスライディングウィンドウを用意し，そのウィンドウごとにその最大値（max pooling），あるいは和（sum pooling）を抽出する処理を実行する．このような処理を行うことにより以下のような効果が期待できる．

・わずかな位置変化に左右されなくなる
・スライディングウィンドウのステップ幅を1よりも大きくすることにより，以降のネットワークのパラメータ数を少なくすることができ，計算コストの低下が見込める．
・過学習の抑制

このような畳み込み層とプーリング層を複数重ね，最後に全結合層を接続する．全結合層において各々のノードは複数の入力に対して以下の計算により出力を得る．

$$x_i' = w_{i1}x_1 + \ldots + w_{in}x_n$$

ここで，xは全結合層の入力となるn次元ベクトル，w_{ij} は全結合層の重みである．この出力対し，さらに活性化関数としてsoftmax関数を用いる．

このようなモデルを構築し，次に学習フェーズにおいてパラメータの学習を行う．すなわち，畳み込み層のフィルタと全結合層でのノード間の重み及びバイアスを訓練データに基づいて計算する．

まず初めにフィルタなどのパラメータをすべてランダムに設定する．次にそのモデルに対して訓練データの入力を与えて，そこから得られた値と正解の値の差の二乗の和を求め（誤算関数），確率的勾配降下法を使って最小二乗法の解を求める．確率的勾配降下法において，多次元空間内である適当な地点において誤差関数を微分し，その傾きが降下している方向にパラメータの値を更新する．以上のような学習方法はバックプロパゲーションと呼ばれている．

以上の処理により，任意の入力に対して，その入力が認識対象のどのクラスに属するかの尤度を出力する．

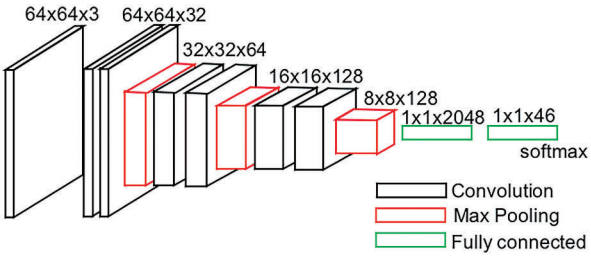

図2. CNNの構造

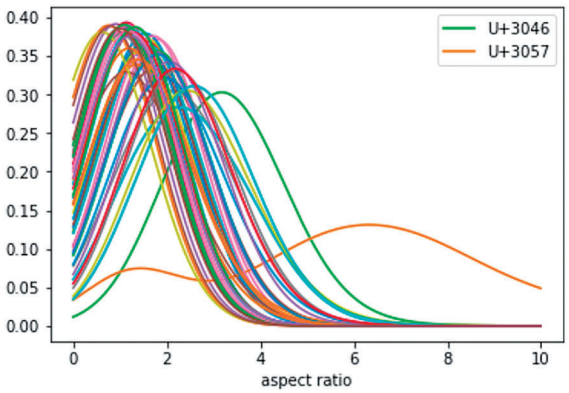

図3. 各文字に対する $p(y|a)$

3.2 文字のアスペクト比を用いた文字認識.

提案手法では，CNNによる文字認識結果に加え，文字のアスペクト比を用いた認識も行い，これら2つの認識結果を統合することで精度向上を目指す．例えば，ひらがなの「し」に相当する文字は，他のひらがなと比較してもアスペクト比が縦に大きく，単にアスペクト比を用いるだけでもある程度の認識ができることが期待できる．

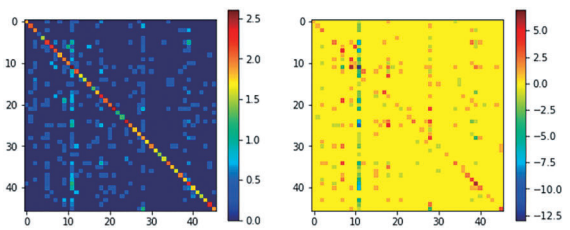

図4. 混同行列

このような知見を利用するため，提案手法では文字のアスペクト比を与えた際の文字の条件付き確率を計算する．ある入力xとその入力のアスペクト比aが与えられた際に，xが文字yである確率は，ベイズ則を用いて次式で表される．

$$p(y|a) \propto p(a|y)p(y)$$

$p(a|y)$については，訓練データから各文字毎にアスペクト比のヒストグラムを算出することで推定することができる．また，$p(y)$についても訓練データから各文字のデータ数を計算することによって得ることができる．

これらより，$p(y|a)$とCNNから得られるxが文字yである尤度（CNN(y|x)）を用いて，xがyである対数尤度を次式で与える．

$$log\ p(y|a,x) = \log CNN(y|x) + \beta \log p(y|a)$$

ここでβは2つの予測結果の重みである．$log\ p(y|a,x)$を最大とするyを認識結果として出力する．

4. 実験結果

学習データとしてくずし文字画像44,353枚を用い，CNNの学習器とアスペクト比を用いた学習器を構築した．図3にアスペクト比aに対する各文字の生起確率p(y|a)を示す．ひらがなの「し」（U+3057）や「う」（U+3046）について，他の文字と比べて大きなアスペクト比（縦長）における確率が高くなっていることが確認できる．

次に，5,000枚のテスト画像を用いて認識を行った結果を表1に示す．CNNのみを用いた際の認識率（89.32%）に比べ，β=2としたときの認識率（90.02）が高く，アスペクト比を併用することによる効果が示された．

また，CNNを用いた際の混同行列を図4左に，β=2としてアスペクト比を併用した際の混同行列とCNNを用いた際の混同行列との差を図4右に示す．なお，可視化のため，混同行列の値は対数スケールで表記している．図4左において11列目（ひらがなの「し」に対応）の非対角成分が比較的大きく，「し」の認識に失敗しているのに対し，アスペクト比を併用することにより，図4右に示すように11列目の値が改善していることが確認できる．しかしながら，その反面，他の文字の認識結果が改悪されており，今後手法の改良が必要である．

表1　認識率の比較

	CNN	β=1	β=2	β=5
認識率	89.32	89.94	90.02	89.10

5. 結論

本研究では，日本語のくずし文字認識を目的として，CNNを用いた識別器とくずし文字のアスペクト比に基づく識別器とを併用した認識手法を提案した．今後の課題としては，単に1文字分の画像を用いるのではなく，前後の文字の情報を利用した精度向上が考えられる．

謝辞

本研究はJSTグローバルサイエンスキャンパスの支援のもとおこなった．ELCAS事務局の方々には様々な方面で支えていただいた．貴重な機会を私に与えてくださった多くの方々にこの場を借りてお礼申し上げる．

参考文献

1. Karen Simonyan, Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. arXiv:1409.1556. (2015).
2. Alex Krizhevsky, et.al. “ImageNet classification with deep convolutional neural networks”, NIPS, 2014.
3. “PRMUアルゴリズムコンテスト2017”，https://sites.google.com/view/alcon2017prmu/

Old Japanese Character Recognition by Convolutional Neural Net and Character Aspect Ratio

Keisuke Ueda[1], Motoharu Sonogashira[2] & Masaaki Iiyama[3]*

[1]Rakunan High School, [2]Graduate School of Informatics, Kyoto University, [2]Academic Center for Computing and Media Studies, Kyoto University

Abstract

Recognizing characters printed on paper or images is important not only for business use but also for literature research. In this paper, we propose a method for recognizing old Japanese characters printed on old documents. Our method combines two classifiers for improving accuracy. One is a convolutional neural net (CNN) that offers good performance for modern character recognition, and the other is a classifier based on the character aspect ratio that can discriminate specific old Japanese characters. Experimental results show that the proposed method achieves better performance compared with the conventional CNN based classifier.

Key words: Character Recognition, Convolutional Neural Net, Aspect Ratio

*Corresponding Researcher: iiyama@mm.media.kyoto-u.ac.jp