値札文字認識による実世界価格比較サイトの実現

成沢　淳史† 　柳井　啓司†

† 電気通信大学情報理工学部総合情報学科　〒 182–8585 東京都調布市調布ケ丘 1-5-1
E-mail: †narusawa-a@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　近年, ウェアラブルデバイスなどから体重や血圧などのロギングを行い, 健康管理やログを分析して効率化を行うライフログと呼ぶ取り組みがある. ライフログとしてウェアラブルカメラなどから画像の形でログが残る場合があり, この画像中には文字が重要な情報として写っている場合がある. 例としてスーパの商品情報が挙げられる. 価格参考のために保存される価格情報はデジタルデータとしてログを残すほうが利用性が高いと考えられる. このため画像中から情景文字認識を行い文字を読み取る必要がある. そこで本研究では商品画像中からの商品名と値段の読み取りの挑戦を行った. 情景文字認識において従来手法とノイズに堅牢な CNN を使った手法を試すことでノイズの多いログ画像から文字の読み取りを行う. 実験ではテスト画像の 48%の値段を読み取れた一方で商品名の読み取りは難しいことが分かった.

キーワード　文字認識, ウェアラブルデバイス

1. はじめに

1.1 背　　景

近年, 健康や日常生活での効率化を求める目的でライフログと呼ばれる取り組みが増えてきている. この取り組みは体重や血圧などをデータとして記録するばかりでなく, デジカメやスマートフォンのカメラで食事を記録するなど幅広い方法で行われている. 今回, このライフログと呼ばれる取り組みとウェアラブルデバイスに注目した. Google glass などのデバイスにはカメラが内臓されており, ライフログとして画像を気軽にログとして残すことができる. このデバイスの利用をライフログの取り組みの範疇で考えた時, 今回はスーパーでのシーンを想定した場合, 商品の価格情報をロギングするのが適切であると考えた. スーパーの価格情報はインターネットでは見られない情報が多い. このため, 普段我々は記憶を頼りに他店との比較を頭の中で行い商品を選ぶことを日常的に行っている. しかし, 記憶が完全なものでないためにしばしば悩むことがある. このため, こうした行動の代替としてライフログとして残した情報からデバイスを通してリコメンドさせたいと思う. しかし, この機能を実現するためにはログ画像から商品と価格情報を自動で認識しなければならない. そのため商品が写る画像に対して情景文字認識の検討を行う.

1.2 目　　的

スーパーの価格情報のように普段, 我々が目にするような文字情報はインターネットでは見られないローカルな情報が多い. そこでライフログと組み合わせることでローカルな情報を検索できるようなシステムを構築したい. 本研究ではこうした応用への第一歩としてまず情景画像中からの文字の読み取りに挑戦する. 現在の情景画像中からの文字認識において, 特に文字検出におけるノイズを減らす研究は現在もなされており困難な課題である. 本研究が扱う画像も例外でなくノイズが複数含まれている. そこでライフログとして記録される画像の中から特に買い物時に写る商品とその値札が写るものを情景画像と想定し, その商品名と値段の読み取りにターゲットを絞りノイズに強い高精度な認識エンジンの構築を目標とする.

2. 関連研究

手書き文字認識や文書読み取りに CNN を使った研究では LeCun ら [1] の研究が有名である. 特に LeNet(図 1) と呼ばれるこのネットワークは 畳み込み層 2 層から成る構成であり, 小規模なデータセットの学習などに利用される. 文字認識は従来からパターン認識の位置づけにあり, 特徴量の設計から評価まで考慮する必要があった. 特に人手で設計された特徴量は輝度の変化や変形に弱いうえ, こうした特徴量をもとにベイズ識別器や SVM などで分類を行うために特徴量を決定し識別器の設計を行うことは難しい作業であった. しかし, CNN は特徴量を畳み込みフィルタの重みとし出力はバイアスと重み計算により求まる結果を正規化するだけで分類を行うことができる. また CNN はプーリングと呼ばれる位置ずれに対する機構を備えており, さらにノイズや輝度の変化に対しては環境の変化をカバーできるような学習データを与えることで対応することができる. こうした理由からシーン文字認識においても CNN は有効であると考えられており, 実際にシーン文字認識への応用が英字において研究され既存の認識手法よりも高い精度を示している. 論文 [2] [3] ではシーン文字認識に必要な学習データをフォントと合成データから作成しデータセットとしており学習データの量とパターンを増やす目的として自動で文字パターンを生成する試みも行われている.

近年, 検出のタスクでは似た領域ごとに分割するセグメンテーションのような処理を行い, 物体が位置するらしい領域に候補を得る Selective-search [4] と呼ばれる手法を用いる. 検出タスクでは Selective-search により大量の領域候補をバウンディングボックスで提供を行い候補からノイズの除去を行っている. 近年, こうした流れからノイズの除去には CNN を用いた R-CNN [5] と

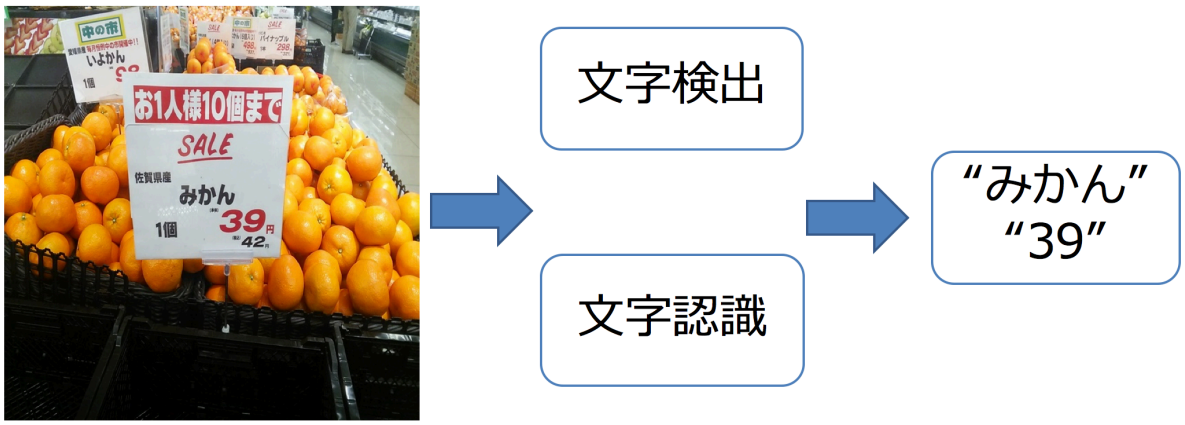

図 2　文字認識の流れ

呼ばれる手法が使われている. また文字検出においてもこうした物体検出と同じ手続きを踏むことで文字の検出を行うことができる. 本研究の文字検出では MSER [6] で提供される大量の候補に対して CNN の Soft-max の出力を用いてノイズの除去を行いながら文字の検出を行う.

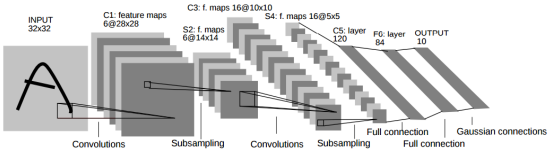

図 1　LeNet(論文 [1] より引用)

3. 手 法 概 要

本研究において情景画像中のシーン文字を読み取り, デジタルデータとするまでの流れを図 2 に示す. 今回は一般のカメラで撮影されたスーパーの商品画像を入力とし, 画像中から商品名とその商品の値段を認識対象とする. 読み取り対象となる商品は野菜などに見られる平仮名, カタカナを中心とした文字を取り扱う. また日本語の読み取りに使う CNN はフォントから学習したものを用いるため情景画像中の商品名と値段がフォントで書かれたものを対象としている.

4. 手 法 詳 細

4.1 本手法における CNN

本研究では検出と認識のそれぞれのタスクで CNN を利用する. CNN の学習にはオープンソースの Deep Learning 用フレームワークの Caffe [7] を利用する. 今研究では LeNet をベースにしたネットワークを利用しする. はじめ, 入力サイズをデフォルトサイズの (32x32) で日本語フォントから学習を行い, シーン文字の読み取りを行うと全く識別することができなかった. そのため, 入力サイズはシーン画像中の文字サイズを想定し (56x56) のグレイスケールとすることにした. また日本語の読み取りでは 100 を超える文字の種類に対応するため LeNet では畳み込み層が 2 層であったところを 1 層増やし 3 層とした. 畳み込み層 (conv) とプーリング層 (pool) の括弧内をそれぞれ (カーネルサイズ x カーネル数, ストライド) とし完全結合層 (ip) の括弧内を出力サイズとして記述すると今回実験に使ったネットワークは図 3 のようになり, input(1x56x56) - conv1(5x5x5, 1) - pool1(5x5,2) - conv2(5x5x10, 1) - pool2(2x2, 2) - conv3(4x4x50, 1) - pool3(2x2, 2) - ip1(100) - ReLu - ip2(122) - Soft-max のように記述される.

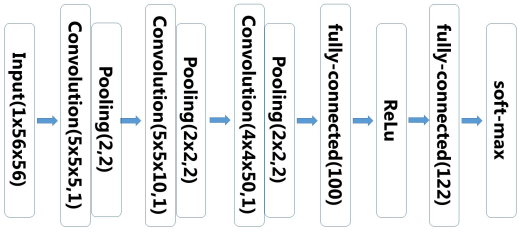

図 3　CNN 概要

本手法では上述のネットワークの最終出力のサイズを変え用途別に表 1 のように 3 種類の CNN を利用する. 文字検出においては通常, 文字と非文字クラスの 2 クラス分類によりノイズの除去を考えるが本手法では検出の段階で商品名と値段のカテゴリ分けを行うため文字, 非文字, 数字の 3 クラス分類可能な CNN を利用する. そのため検出用途での CNN は値札から文字を切り出し各クラスそれぞれ 1062 枚, 797 枚, 574 枚とし学習を行った. また文字の認識では商品名を読み取る CNN と値段を読み取る CNN を 2 種類用意した. 商品名の読み取りには文字の種類が多い日本語への対応としてフォント画像 4 から学習した CNN を利用する. 日本語フォントの描画には SDL ttf ライブラリ (注1) を用いて OpenGL 上でスケールを 0.9 倍から 1.2 倍, 30 度ごとに回転を与えたデータを出力し各クラス 700 枚ほどで用意した. 今回, ひらがな, カタカナ, 数字に加え簡単な漢字を含めた 122 文字に非文字クラスを加え日本語用の文字識別用 CNN としている. また値段の読み取りには実データを利用する

(注1) : 'https://www.libsdl.org/projects/SDL_ttf/'

図 4　フォント学習データ

ことで精度向上を図る. このため数字 10 クラスのデータを実データから用意しデータ量が十分でないクラスにはフォント画像から補い各クラス 100 枚程度になるようデータセットを作成し CNN を学習した.

	検出用	認識用 (日本語)	認識用 (数字)
学習データ	実データ	フォント	実データ+フォント
認識クラス	3	122+1	10
学習枚数 (枚)	2433	各 700	各 100

表 1　用途別 CNN の学習データ

4.2　シーン文字検出

文字候補の検出には MSER [6] を利用する. この MSER を利用し図 5 のように大量の文字候補バウンディングボックスを得る. 今回 CNN への入力サイズのアスペクト比が 1:1 であるため提供されるバウンディングボックスは長い辺に合わせて修正している.

図 5　MSER により得られた BB

次にこの大量の文字候補バウンディングボックスから文字のみが残るようノイズの除去を行う. 文字 C_1, 非文字 C_2, 数字 C_3 の 3 クラスへ識別可能な CNN の Soft-max 出力 $P(C_i)$ を用いて図 5 から文字 C_1 と数字 C_3 が残るようノイズの除去を行う. ここでは Soft-max 後の正規化出力の非文字クラスである確率が

$$P(C_1) + P(C_3) < P(C_2) \tag{1}$$

である時に,　文字候補を誤検出として除去し文字か数字である場合には CNN の正規化出力 $P(C_1)$ と $P(C_3)$ の大小関係から数字と文字の分類を行い図 6 のように文字と数字でのカテゴリ分類に利用する.

図 6　ノイズ除去後 (左:文字 右:数字)

ノイズ除去後, 残った候補を画像上の X 軸の左から順に並べ隣合う文字を探索し連結する. 隣合う文字の探索にはテキストの文字が水平方向に並んでいると考えて最も近傍にある文字候補に連結する. この時, 単純な x 軸でのソートでは上下に位置する文字候補へ連結してしまうため式 2 のように y 軸に対して重みをつけたユークリッド距離 d が最小となる文字候補に連結する. また $d < 100$ である場合には連結を止めそれまで繋がった文字候補を一つのテキストまたは単語としてグループ化を行い, さらに文字がグループに 1 つしかない場合にはノイズとして除去を行う. ここで実験的に α の値は 5 に設定した.

$$d = \sqrt{(x_1 - x_2)^2 + \alpha(y_1 - y_2)^2} \tag{2}$$

またテキストラインに並ぶ文字はある程度同じサイズにあると考えられる. このためテキストライン上の文字候補のサイズに関する分散を求めこの分散が極端に大きい場合には文字の列ではない可能性が高いためこのグループを除去している. この一連の処理により文字と数字のそれぞれに対して図 7, 図 8 のような結果を得る. そして, このグループの文字を 1 文字づつ認識することで商品画像から値段と商品名の文字列の読み取りを行う.

図 7　文 字 (BB)

図 8　数 字 (BB)

5.　実　　験

実験ではフォントから学習した CNN 識別器の認識評価を行い, 次に情景画像中からのシーン文字検出から認識までどの程度読み取ることができるかテスト画像を 62 枚用意し実験した.

5.1 文字認識実験

フォントのみから学習した CNN を利用し, 身近なシーン文字を正しく読み取りできるのか実験を行った. 実験に利用する CNN 識別器はひらがな, カタカナ, 英字, 数字の 122 文字と非文字を認識できるよう学習している. 読み取り結果が良かった例から結果を載せる. 図 9 のような商品の値札に対して, 図中から手動で切り出した文字画像を入力とし認識を行った. この図中の各文字の予測結果の上位 3 位までは表 2 のような結果となり, 商品名と値段の読み取りが行えることがわかった. また読み取りに失敗した例 (図 10) を載せる. 値段の読み取りが出来ているが, 商品名の読み取りは失敗している.

図 9 “れんこん” 商品札

	れ	ん	こ	ん	9	8
1 位	れ	ん	こ	ん	9	8
2 位	カ	へ	ニ	ル	ヨ	ト
3 位	ヤ	へ	さ	イ	タ	タ

表 2 “れんこん” 認識結果

図 10 “パプリカ” 商品札

	パ	プ	リ	カ	1	2	8
1 位	1	フ	リ	ガ	1	2	8
2 位	パ	プ	?	つ	?	里	ヌ
3 位	バ	ラ	ツ	フ	ア	さ	ミ

表 3 “パプリカ” 認識結果

また読み取りの対象となる値札の画像を集め, 画像中から文字を切り出し予測結果の 1 位 (@1) から 5 位 (@5) 以内に正しい文字が現れる回数をカウントし分類率を求めた. 収集した文字 398 文字のうち 342 文字が識別可能な文字であり, 認識結果は表 4 のようになった, 1 位での分類率は 46.8%, 5 位までの結果は 70.8%であった.

上位 (位)	@1	@2	@3	@4	@5
分類率 (%)	46.8	58.8	66.9	68.4	70.8

表 4 分 類 率

5.2 値段認識実験

値札が写るテスト画像を用意し, 実際に文字の検出から認識までどの程度行えるのかを実験した. 図のような画像とテキストを出力し, テスト画像 62 枚を評価し集計を行った. 本研究の目的では商品名の読み取りまで視野に入れているが, 文字認識実験より文字の予測結果 1 位からの商品名の読み取りは難しいことがわかったため, 今回は値段のみ評価を行った. 評価には入力画像に値札が写るものとし, 値段が検出された領域の値段の読み取りに成功した画像がテスト画像 62 枚のうち何枚あるかをカウントした. その結果, 検出から値段の認識まで成功したものは全体の 48% であった. 検出から認識までにかかる時間は文字検出後の候補数に依り, およそ 5 秒以内となった. 正しく値段の検出と認識が行えた例 (図 11) と出来なかった例 (図 12) をそれぞれ図に示す. 数字に関しては “2”, “3”, “8”, “9” といった数字の読み取りはある程度正しくできる一方で “1” の数字の検出が難しいことが図より分かる. また照明が暗い場合に検出が難しいことがわかった.

6. 考 察

6.1 文字認識精度検討

フォントのみから学習した CNN で文字認識テストを行った結果ではトップ 1 での認識率では 46.8% の結果を得た. より高い精度を目指す場合, まずフォントのみから構築されたデータセットでは実世界のシーンをカバーするのに十分な歪みや背景を再現出来ていないためノイズを考慮した背景を複数用意しておき, 輝度変化へは学習画像にランダムな濃淡を与えるなどの方法が必要と考えられる. またスケールの問題が考えられる. MNIST は入力が 28x28 であるが日本語に対してこのサイズで CNN の学習を行い今回のような実験を行ったところ全く文字を識別することが出来なかった. このため入力サイズを実際の撮影写真から想定し 56x56 として CNN を学習している. このことから日本語の場合にはある程度のサイズを保証して文字画像を用意しなければならないことが分かった. しかし今回の実験では文字が小さい場合にはリサイズにより 56x56 としていたため実験のレギュレーションに問題があったと考えられる. このため認識対象の文字のサイズをある程度保証することで今回の実験結果よりも良くなるだろうと考えられる. 今回, 日本語 122 文字を識別するよう CNN を用意した. 詳細な実験を行っていないが他クラス分類として識別可能な文字の数を増やして行くと徐々に文字の識別が難しくなるようである. この原因として特に文字においては文字自体のパターンの特徴が少ないことが原因と考えられる. 例として “フ” と “ラ” では同じ位置に特徴的な線が見受けられるが, こうした共通の特徴に CNN の畳み込みフィルタが反応を示すために識別が難しい一因になっている可能性がある. 人間でさえ “1” と “7” を間違えることが多く CNN の畳み込みフィルタにもこの直線成分に反応が

図 11　値段の読み取り (成功例)

図 12　値段の読み取り (失敗例)

見受けられる. このように文字は人工的なものであるとともに直線と簡単な曲線から構成されるために文字の種類が豊富な日本語においてはいくつか似ている文字が存在する. このため表 4 の結果ではトップ 1 では 5 割以下の分類率がトップ 5 では 7 割となっている. また表 2, 表 3 からも文字の予測結果の上位には人目からも特徴が似通っていると思われる文字が現れる結果となっている. このため予測結果トップ 1 を使っての商品名の読み取りは難しいものであり, 日本語の商品名の読み取りには予め辞書を用意しておき, 最もそれらしい単語とのマッチングを行うものが理想と考えられる.

6.2　シーン文字認識フレームワーク評価

実験では全体の 48% 程度のテスト画像から正しく値段を読み取ることができた. 実験を通して条件がよいシーン画像に関しては値段の読み取りが行えている印象を感じた. 一方で視点や照明の条件が悪い場合には検出, 認識ともに難しい結果となった. 視点や照明条件をカバーできるよう更にデータセットを拡充することで結果の向上が期待できると考えている. また数字の検出に関して図 12 に見られるように “1” の検出が難しいようである. これはノイズとして検出された領域が CNN の畳み込みフィルタには縦の直線として反応することが原因と考えられる. 値札のエッジを始めとして値札を支える棒や柵など直線的なデザインが多い人工物が映り込む場合, “1” に類似し誤検出となるため負例としてデータセットに加えていくことを繰り返しノイズ除去の精度を上げる取り組みを行ったため “1” が負例になる確率が高くなった結果と考えられる.

また残念ながら商品名の読み取りに関しては課題を残す結果 (図 13) となった. 検出時において一部の文字が欠落し, 認識においも誤認識が多く実用的ではなかった. MSER の候補検出において “こ” など全体を囲えない文字も見られ課題を残す結果となった. 文字列の領域までは図 13 のように分かる場合もあるためこの領域に対して Sliding Window 法を用いるなどよりセンシティブな検出を検討する必要があると考えられる.

7.　おわりに

7.1　まとめ

近年の機械学習手法としてディープラーニングを行うことができるフレームワークが数多く登場し物体認識が広く応用されている. 本研究のシーン文字認識タスクにおいても CNN は多クラス分類への認識利用のみならず, R-CNN として検出への利用が可能であった. 結果として画像認識を組み込んだアプリケーション開発において想定される検出から認識までの一連の工程を実現することができた. 今回の実験では商品札から値段の読

図 13　商品札の読み取り

み取りに関してはテスト画像の 48%を認識する結果となった. また実験結果から商品名の読み取りは難しいことがわかった. 実験を通して商品札の文字を読み取るのに実用的な CNN のモデルを学習させることは難しく, 実現のためには用途に見合った学習画像を収集する必要があると感じた. 既存のデータセットや Web の大量の画像をマイニングし利用することで画像を集めることができるが, 特定の用途に見合ったシーンを想定した画像のみを集めることは難しいため地道に実画像を撮影する必要があり, データセットの作成は難しいタスクであると感じた. 現状, データセットの作成にはクラウドソーシングに依るところをみると, 個人で行える範囲でデータセットを作成する方法としてデータを自動生成する技術が有望に思えるが, 実データをカバーできるような画像の生成は未だ難しいと考えられる. このためデータセットの構築を効率よく行う方法を模索する方が現状良いと考えられる. 今研究のため値札が写るスーパーの商品画像を収集した. こうした画像は Web 上からの収集は難しいためにオートマチックに集めることが出来ない. このためこうした新規のデータからサービスを展開することはコストが必要となり新規性のある応用への期待が薄いと考えられる. しかし今後より幅広い分野へ応用を考える場合, 目的を達成するのに必要なデータセットを増やすことがまず第一の課題であると考えられ本研究においてもまた同様の課題であると考えられる.

7.2　今後の課題

情景画像中にはシーン文字の他に物体や背景が写っており領域を分割することができる. 現在, こうした領域のコンテキストと文字認識を組み合わせることで誤検出の削減や認識精度の向上が図られているようである. 本研究のターゲットにおいても例外でなく, 値札と商品とで領域分割を行うことが可能である. 値札の領域に対し限定的に読み取りを行うことで誤検出を減らすことができるうえよりセンシティブな処理を行うことができると考えられる. また商品自体を物体認識することで, 商品名の読み取り結果を訂正, 修正することができる. 「長崎県産トマト」などは「トマト」のようにシンプルな認識結果を期待することができる. また今回はスーパーの商品を認識対象として絞り込み文字の種類を限定することで文字の出現頻度の調査が可能であると考えられる. スーパーに見られる文字は大体何種類あるのかを知るばかりでなく単語帳の作成や文字の出現頻度を調べることで今回の文字認識の結果に組み合わせることができると考えている. 特に商品名は辞書を予め用意しておき, 文字認識結果をその辞書の中から探し最もそれらしい商品名を返す手段が有効であると考えられるため今後調査を行いたい.

文　　献

[1] L. Yann, B. Léon, B. Yoshua, and H. Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[2] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *Proc. of NIPS Workshop on Deep Learning*, 2014.

[3] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *Proc.of European Conference on Computer Vision*, 2014.

[4] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, Vol. 22, No. 10, pp. 761–767, 2004.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.