講演番号：133

チラシ画像からの商品情報自動抽出－精度改善－

D－11　Automatic extraction of goods information from pictures of leaflets
－Accuracy improvement－

柴山　美沙希
Misaki Shibayama
芝浦工業大学
Shibaura Institute of Technology

高橋　正信
Masanobu Takahashi
システム理工学部
College of Systems Engineering and Science

1. はじめに

チラシの情報をデータベース化できれば商品の旬や最安値が分かり便利だが，チラシ情報はテキストデータになっておらず，OCR ソフトでもほとんど認識できない．チラシを収集し人手でデータ化するサービス[1]もあるが，企業向けで高額である．そこで，チラシ画像から商品情報を自動認識してデータベース化する機能の実現を目指す．商品情報とは，価格と価格以外の内容情報(会社名または産地，商品名，内容量)のことを指す．これまで埼玉県に多く店舗のあるヤオコーのチラシを対象として，価格と税込か税抜かを自動認識する機能[2]を実現した．内容情報については Google Cloud Vision API を利用して自動認識する機能[3]を実現したが，精度が不十分であった．そこで本研究でその改善，特に商品名認識の精度改善を図った．

2. 内容情報認識機能

Vision API に内容情報を含む画像を入力として与えると，1 文字毎の認識結果(テキストデータ)と位置情報(外接長方形)が得られる．そこで，位置情報を元に同じ行の文字を結合する．さらに上下に近接する行を結合してブロック化し(図 1)，ブロック内の文字を一つの文字列とする．

(a) 行内の塊を囲んだ画像　(b) ブロックを囲んだ画像
図 1 複数行のブロック化

文字列は会社名(産地)，商品名，内容量からなるため，次にそれらの単語を分離する．まず，内容量は文字列の最後にあり，「数字」+「単位」で構成されるため単語を特定して抽出する．残る会社名(産地)と商品名の分離には文字の高さを利用する．通常，会社名よりも商品名の文字サイズが大きいため，k-means 法を用いて文字の高いクラスと低いクラスに分け，高いクラスを商品名とする．

作成した内容情報には，Vision API の誤認識や単語の誤分離により誤字脱字が含まれる場合がある．本研究では，事前に作成した商品情報データベースを利用することで，商品名の誤字脱字を修正する機能を実現した．

3. MySQL

商品情報データベースには MySQL を用いた．MySQL とはオープンソースで公開されているリレーショナルデータベース管理システム(以下「RDBMS」)の 1 つである．Windows を含む多くの OS で利用することができ，日本語にも対応している．MySQL を始めとした多くの RDBMS では全文一致検索の他，ワイルドカードを用いたあいまい検索が可能である．使用者が文字コードや照合順序を設定することができ，日本語の場合，濁音，半濁音の区別や平仮名，片仮名の区別の切り替えが可能である．

4. 修正手順

初めに対象の内容情報を全文一致で検索し，一致しなかったものに対して図 2 に示すように組み合わせをずらしながら 2 文字ずつあいまい検索を行う．このとき Vision API において濁点，半濁点が誤認識される場合があるため，濁音，半濁音の区別はつけない．検索文字が含まれる全ての単語と各単語の検出回数を求め，検出回数が最多のものを採用する．回数が同値のものが複数あれば，文字数が元の内容情報に近い単語を採用する．

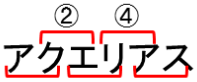

図 2 検索文字

表 1 検索結果

検出回数	検出単語
5	アクエリアス
5	アクエリアス　ゼロ
⋮	⋮

5. 実験

ヤオコーHP からダウンロードしたチラシ(2018/04/21 号)[4]を用いて実験を行った．データベースはヤオコーの通販ページ[5]から取得した約 9000 件の商品情報を用いて作成した．誤字脱字修正機能の評価のみを行うため，単語の分離は手動で行った．作成したデータベースに商品情報が存在する 104 個の商品名に対して実験を行った結果を表 2 に示す．Vision API で誤認識された商品名 13 個は全て正しく修正された．

一方で，元々正しく認識された商品名が誤った単語に変換される場合があった．その原因はチラシの表記とデータベースの表記の相違である．チラシでは商品名「つき」(表 3)の商品がデータベースでは「つき　パック」(表 4)と登録されているため，「みかづき」と誤修正された．この問題はチラシの表記に沿ったデータを登録することで解決できると考える．また，今回は商品名のみを修正したが，会社名も同時に検索することでも改善できると考える．

表 2 商品名認識結果

修正前	正		誤	
	91		13	
修正後	正	誤	正	誤
	86	5	13	0

表 3 認識結果(修正前)

会社名・産地	商品名	内容量
月桂冠	つき	2l1本

表 4 検索結果

会社名・産地	商品名	内容量
おやつのくに	みかづき	72g
月桂冠	つき　パック	2l

データベースを用いることで Vision API の誤認識による商品名の誤字脱字を修正することができた．今後は単語の分離方法の改善などにより，内容情報全体の認識精度を向上し，実用的な機能を実現したい．

[参考文献]

[1] 株式会社ドゥ・ハウス，全国チラシ情報サービスセンター，https://www.chirashiinfo.jp/, 2018 年 12 月 20 日閲覧.
[2] 染谷謙太朗，他：信学会東京支部学生会研究発表会，154, 2014.
[3] 石井宏樹，他：信学会東京支部学生会研究発表会，156, 2017.
[4] ヤオコー，https://www.yaoko-net.com/, 2018 年 4 月 21 日閲覧.
[5] YAOKO ネットスーパー，https://www.ns.yaoko-net.com/front/app/common/index, 2019 年 1 月 9 日閲覧.