チラシ画像の商品情報認識―高機能化と自動化―

D－11

Automatic extraction of goods information from pictures of leaflet
－Functional improvement and automation－

原田　靖之
Yasuyuki HARADA
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. 背景

チラシに掲載された商品情報(会社名，商品名，内容量，価格)を自動でデータベース化して比較できる機能があれば，最安値や販売の傾向などが一目で分かり，消費者にとって便利である．しかし，チラシ情報はテキストデータになっておらず，また，価格に特殊なフォントが多用されるため，既存の OCR ソフトを用いた文字認識精度は低い．全国からチラシを収集しデータ化するサービス[1][2]もあるが，企業向けで高額である．我々が調べた限り，チラシ画像中の商品情報を自動認識する機能を実現した研究は無い．なお，商品情報は内容情報(会社名，商品名，内容量)と価格から構成される．

我々は，チラシ画像から価格情報を自動認識する機能[3]，内容情報を自動認識する機能[4]，価格情報と内容情報を対応付ける機能[5]を実現した．しかし，内容量の認識精度の改善と，箇条書き表記の商品名への対応，および処理の自動化が課題となっていた．

2. 目的

本研究の目的は，内容量の認識精度の改善，箇条書き表記への対応，および商品情報の認識結果の出力までを全自動で行う機能の実現である．なお，本研究では埼玉県に多く店舗のあるヤオコーのチラシを対象とした．

3. 手法

3.1 内容量の誤認識修正

グラム数や個数などの内容量はフォントが小さいこともあり本システムが文字認識に利用する Google Vision API[6]で正しく認識されない場合が多かった．そこで，本システムで作成する商品情報のデータベースを利用し，以下の手法で修正することで精度改善を試みた．

① 認識結果がデータベースの内容量と一致する場合：修正を行わない
② 認識結果がデータベースの内容量の後ろに文字が追加された文字列の場合：データベースの情報に修正
③ それ以外(誤認識の場合)：価格の認識結果とデータベースに格納されている価格を比較し，最も近い価格の内容量に修正

3.2 箇条書きへの対応

ヤオコーを含む多くのスーパーのチラシでは，複数の商品名を含む箇条書きの表記は，図 1 の赤線で囲った部分のように商品名の前に「●」がついている．そこで，商品名の前に「●」がある場合は「●」の前後で商品名を分割することにより，複数の商品名を認識する．

図 1　箇条書きの表記例[7]

3.3 処理の自動化

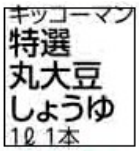
従来システムでは，価格情報の自動認識，内容情報の自動認識，価格情報と内容情報の対応付けはそれぞれ別のシステムで動作していた．また，利用するデータベースに含まれるデータも一部異なっていた．そこで，データベースを統合するとともに，必要な情報をファイルで受け渡しすることでシステムを順番に実行できるように改善し，全処理の自動化を実現した(図 2).

図 2　処理の流れ

4. 実験

4.1 内容量の誤認識修正

ヤオコーのホームページ[7]から入手した 12 枚のチラシ画像(表と裏は別のチラシ画像とする)を実験に用いた．チラシ画像中の対象となる内容情報は 145 個である．内容量の認識精度改善前は 145 個の内容情報のうち，会社名，商品名，内容量を全て正しく認識できたものは 79 個(正解率 54.5%)であった．3.1 の手法で内容量の認識結果を修正した場合は，正しく認識できたものは 139 個(正解率 93.1%)となり，正解率を大幅に改善することができた．内容量の修正例を図 3 に示す．



内容量の認識結果：121 本
⇩
修正後の認識結果：1ℓ1 本

図 3　内容量の修正例

4.2 箇条書きへの対応

実験に使用したチラシ画像中で箇条書きとなっている箇所は 40 カ所あり，商品名は 93 個存在する．箇条書きへの対応を行う前は複数の商品名がある場合に 1 つの商品名しか認識できなかったため，93 個のうち会社名，商品名，内容量を全て正しく認識できたものは 39 個(正解率 41.9%)と精度が低かった．3.2 の箇条書きへの対応処理を追加することで，正しく認識できた商品名は 57 個(正解率 61.3%)に改善された．

精度は改善されたもののさらなる改善が必要である．商品名が認識されなかった主な原因は商品名の前にある「●」が Google Vision API により十分に認識されなかったことである．実験では半分程度が認識されなかった．例えば，パターンマッチングなど別手法で「●」の認識精度を改善することで箇条書きの商品名の認識精度をさらに改善できると考える．

5. まとめ

内容量の認識精度を改善し，全処理の自動化を実現した．今後の課題としては，箇条書きの商品名の認識精度改善と，他店のチラシへ適用できるか検証することが挙げられる．

[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.dohouse.co.jp/research/tento_02/>, 2021 年 6 月 20 日アクセス.
[2] 株式会社ナビット，チラシ収集サービス，<https://www.navit-j.com/service/chirashi.html>, 2021 年 6 月 20 日アクセス.
[3] 柴山美沙希，他：平成 30 年度電子情報通信学会東京支部学生会研究発表会，133, 2019.
[4] 柴山美沙希，他：パーソナルコンピュータ利用技術学会論文誌，15 巻, 1 号, pp.32-40, 2021.
[5] 外山壮大，他：令和 2 年度電子情報通信学会東京支部学生会研究発表会，81, 2021.
[6] Google Cloud, Cloud Vision API, <https://cloud.google.com/vision?hl=ja>, 2022 年 1 月 13 日アクセス.
[7] ヤオコー，2021 年 6 月 19 日広告.