

# チラシ画像の商品情報自動認識 – 箇条書きの印認識 –

Automatic extraction of goods information from pictures of leaflets

–Extraction of marks of bullet points–

中山 翔

Kakeru NAKAYAMA

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

チラシに掲載されている情報を自動でデータベース化するシステムがあれば、消費者は一目で最安値や販売傾向を把握することができて便利である。また他店の情報を容易に把握できるようになれば店側も便利になる。しかし、店舗が作るチラシは画像になっていて、更に背景が複雑なため既存の OCR ソフトによる文字認識精度は低い。全国のチラシを収集しデータ化するサービス[1]、[2]もあるが、企業向けかつ高額である。我々が調べた限りでは、商品情報と価格を同時に自動抽出する研究は無い。

我々は商品の内容情報(会社名、商品名、内容量)を認識する機能を実現し、ヤオコーのチラシにおいて認識率 92.9%を実現した[3]。一方でカスミのチラシでは 73.7%と低く[4]、他店舗の認識精度の改善が課題となっている。特に箇条書きの印の認識率が低いことが内容情報認識の精度低下の要因となっている。

チラシでは、1つの価格に会社名、商品名、内容量が1つずつ対応する場合と、複数の商品、内容量、会社名が対応する場合(図1)があり、認識処理を切り替える必要がある。後者の場合は箇条書きで記載されるため、箇条書きの印で判別し、印を用いて認識する。そのため、印認識率が低いと精度が大幅に低下する。そこで本研究では、カスミのチラシにおける箇条書きの印の認識精度の改善を目的とした。

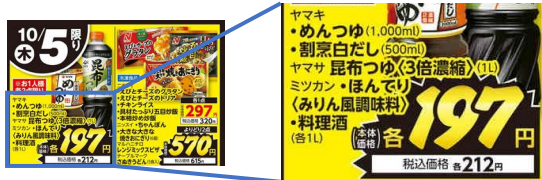


図1 箇条書きのチラシ[5]

## 2. 手法

### 2.1 従来手法

箇条書きの印の認識にはテンプレートマッチングを利用する。複数サイズの「●」のテンプレートを用意し、正規化相互相関が 0.8 以上のものを印の候補として抽出する(図 2(a))。次に、抽出された候補を固定閾値で二値化し(図 2(b))、円形度が 0.8 以上のものを印とする。

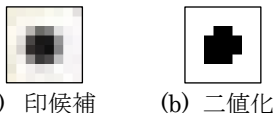


図2 印認識の流れ

### 2.2 改良手法

従来研究では手法やパラメータがヤオコーのチラシに最適化されていたため、カスミのチラシの印をほとんど認識できなかった。本研究ではカスミのチラシに対応するため以下の手法を導入した。

①図3のようにカスミの印はヤオコーの印より小さいため、サイズの小さなテンプレート(黒丸の直径

4,5,6,7,8画素)を用いた。

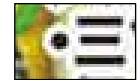
②円形度の閾値を 0.8 から 0.75 に下げた。

③サイズの小さい印ではテンプレートマッチングの誤差が出やすいため、サイズが 4,5,6 画素の場合のみ正規化相互相関の閾値を 0.7 に下げた。

④黒色でない印を除外するため、二値化の前に印の中心 1/3 四方の領域で画素値の最小値を求め、最小値が 50 より大きい場合を除外した。



(a) ヤオコーの印[6]



(b) カスミの印[5]

図3 各店舗の印の比較

## 3. 実験

カスミのホームページに掲載されていたチラシ 17 枚(印 102 個)を用いて実験した。結果を表 1 に示す。従来は 12.7%しか認識できなかったが、①~④の手法を全て追加することで再現率を 93.1%に改善できた。④を無効にした場合は 94.1%となったが、適合率と F 値が低下した。

認識に失敗した印の大半はテンプレートマッチングの正規化相互相関が 0.8 未満だった。チラシ画像は圧縮されているため、圧縮ノイズにより形状が正円から変化したことが原因である。誤抽出としては、商品名や内容量の中黒の誤抽出が多かった。こうした誤抽出に対しては、前後の文字列を利用して印か中黒か判別するなどの対策が考えられる。

今後の課題としては、さらなる精度改善、白色の印への対応、より多くの店舗で汎用的に使えるシステムにすることが挙げられる。

表1 実験結果

	再現率	適合率	F 値
従来	0.127	0.722	0.217
①~④有効	<b>0.931</b>	<b>0.798</b>	<b>0.860</b>
①無効	0.098	0.667	0.171
②無効	0.775	0.814	0.794
③無効	0.873	0.840	0.856
④無効	0.941	0.762	0.842

### [参考文献]

- [1] “全国チラシ情報サービスセンター | チラシ分析、チラシ収集、価格動向 | 株式会社エクスクリエ”, <[http://www.excrie.co.jp/research/tento\\_02/](http://www.excrie.co.jp/research/tento_02/)> 2024 年 1 月 6 日アクセス。
- [2] “チラシ・新聞収集サービス | 株式会社ナビット”, <<https://www.navit-j.com/service/chirashi.html>> 2024 年 1 月 6 日アクセス。
- [3] 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.
- [4] 石川敦斗, 他: 電子情報通信学会東京支部学生会研究発表会, 89, 2023.
- [5] カスミ, 2023 年 10 月 1 日広告。
- [6] ヤオコー, 2023 年 12 月 8 日広告。