講演番号： 188

チラシ画像からの商品情報自動抽出－商品名認識－

Automatic extraction of goods information from pictures of leaflets
－Recognition of Product name－

D－11

亀山　綾乃
Ayano Kameyama
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

折込チラシの情報をデータベース化できれば，商品の旬や最安値などが分かり便利であるが，チラシ情報はテキストデータとして公開されておらず，また 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名と価格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．このうち，チラシ画像から抽出された数字が価格であるか否かの判別機能，チラシ画像から価格情報を抽出できる自動認識機能（認識成功率 99.35%），認識された価格が税込か税抜かの識別機能を実現した[2]（以下従来研究）．しかし，商品名を認識し，更に価格情報と組み合わせる機能は実現されておらず，その実現が課題となるが，今回は商品名の認識機能の実現を図った．なお，この機能は会社毎に実現するが，まずは埼玉県に多く店舗のあるヤオコーを対象とした．

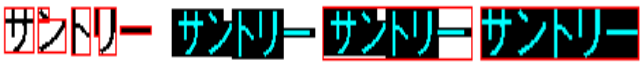
2. 文字候補領域の抽出

最初に，商品名を構成する文字の候補（文字候補領域）を抽出する．手法としては，従来手法の「円領域の抽出機能」で実現した色と領域の形状を利用する手法を改修して用いた．抽出した文字候補領域には価格領域も含まれている．そこで，価格領域を先に抽出し，その領域を文字候補領域から削除した．なお，税込価格については数字以外の「＜税込　＞」の部分が価格領域に含まれないため削除されない．そこで，税込と判別された場合だけ価格領域の左右を大きめに削除し，不要な領域を削除した．

3. 文字候補領域から商品名候補領域を作成

文字候補領域をグルーピングすることで商品名候補領域を抽出する．具体的な手順は以下のとおりである（図 1）．

(a) 文字候補領域のラベルの外接長方形を取得．
(b) 外接長方形を黒画素で塗りつぶしラベリングを行い，各領域を左右と上下に拡大．商品名は横長と想定しているため，横方向への拡大が大きくなるように，パラメータを設定．
(c) 拡大して結合したラベルの外接長方形を取得．
(d) 領域を拡大した分だけ縮小し，外接長方形（商品名候補領域）を取得．



(a)　(b)　(c)　(d)

図1　商品名候補領域作成の処理手順

なお，同じ商品名の文字候補領域は基本的に図 2(a)のように同一色で囲まれている．そこで，文字候補領域の周辺の色を抽出し（図 2(b)），他と異なる色を持つ領域がグルーピングされないようにして誤抽出を低減した．

(a)原画像　(b)領域周辺の色抽出画像

図 2　ラベル周囲の色情報の例

また，商品名候補領域の抽出結果を調べたところ，図 3 右のように数字0の内部など閉じた領域が誤抽出されることがわかった．そこで，ある色で抽出された商品名候補領域が別の色の商品名候補領域の内部に完全に含まれる場合，含まれる方の領域を誤抽出として削除した．

図 3　同部分の比較（左から原画像，黒抽出，白抽出）

4. 実験

ヤオコー公式 HP からダウンロードしたチラシ画像（2014/06/04 号）を 1 面分用いて商品名の抽出と認識の実験を行った．商品名領域の数は 269 個である．抽出実験の結果，再現率 81.0%に対して適合率が 42.4%と低くなった．特に文字色が白の場合の精度が低かった．その原因の一つは，文字色が白の場合，周囲の色がバラバラであるため，商品名候補領域の選別が黒色ほど有効に機能しなかったことである．他の原因としては，色抽出の際に商品写真から誤抽出される領域が多いことが挙げられ，そうした誤抽出の低減が今後の課題となる．次に，商品名候補領域を Adobe Acrobat X の OCR 機能を利用して認識した．その結果，正しく抽出された 218 個の商品名候補領域のうち，商品名が正しく認識されたのは 116 個（53.2%）であった．誤認識が多かったものは，濁点や半濁点，容量の ml やg等の小文字，片仮名の「リ」「ル」などである．

5. おわりに

約 8 割の再現率で商品名候補領域を抽出する手法を実現した．今後は誤抽出の低減による精度の改善とともに，文字認識精度の改善も必要である．例えば，Adobe Acrobat X が認識しやすいように商品名候補領域を加工することなどの改善策が考えられる．

[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.chirashiinfo.jp/.
[2] 染谷謙太朗，高橋正信：“チラシ画像からの商品情報自動抽出－価格認識－”，電子情報通信学会総合大会学生ポスターセッション，ISS-SP-250，2015 .