チラシ画像からの商品情報自動抽出－価格認識－

染谷　謙太朗　　高橋　正信
芝浦工業大学　システム理工学部　電子情報システム学科

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．しかし，店舗側は過去の値段や他店の値段と容易に比較されるのを避けたいため，チラシの情報はテキストデータにはなっていない．また，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考えた．そうした機能を実現した研究は著者らが調べた限り無かった．実現が難しい理由として，背景が複雑でかつ特殊なフォント（文字どうしが重なりあうような「くいこみ文字」）が多用されていることが挙げられる．このため，既存の OCR ソフトを用いてもほとんど認識できない．また，会社ごとのフォーマットの違いも大きい．そこで，認識機能を会社ごとに実現することとした．今回は埼玉県に多く店舗のあるヤオコーを対象とし，まず「価格」の認識機能の実現を図った．

2. 数字／円領域の抽出・認識

ヤオコーのチラシの価格に必ず付いている漢字の「円」の領域（以下「円領域」）と数字領域を抽出する．チラシ画像内で価格の表記に使われている 4 色（赤，黄，黒，白）のそれぞれについて色の条件を用いて2値化し，面積・幅・高さ・縦横比などの条件を満たしていない領域を削除する．

次に，抽出した数字／円領域を認識する．チラシはフォントが限られる／特殊なフォント（くいこみ文字）が利用される／パラメータの調整が可能などの理由から，認識手法としてテンプレートマッチングを用いた．フォントの種類ごとに求めたサイズ分布をもとに，数字についてはゴシック体 8 サイズ，くいこみ文字 13 サイズ，円については傾き無し 2 サイズ，左傾き 3 サイズ，右傾き 1 サイズのテンプレートを作成した．高さの近いテンプレートに認識対象の領域を変形させた後マッチングして認識するが，シミュレーションの結果，数字領域には残差，円領域には相互相関係数を用いることとした．数字あるいは円で無い領域は閾値を設けて除外する．

3. 数字領域と円領域の結合

隣接する数字領域と円領域を結合し，価格と認識する．具体的には，全ての円領域について以下の処理を行う．

(A)円領域と数字領域の結合：円領域の左端の座標から左へ探索し，距離d_1以内に下端の位置の差が一定範囲以内の数字領域が見つかれば結合して(B)へ．見つからなければ終了．

(B)数字領域同士の結合：結合された数字領域の左端の座標から左へ探索し，距離d_2以内に下端の位置の差と高さの差が一定範囲以内の数字領域が見つかれば結合．この処理を数字領域が見つからなくなるまで反復．

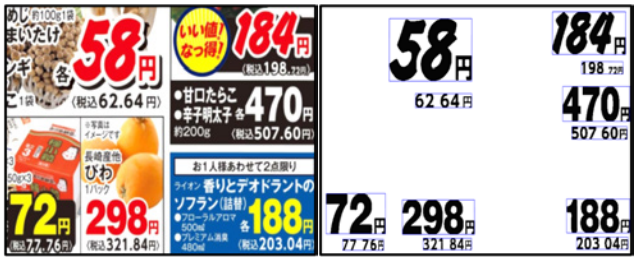
4. 税込／税抜価格の識別

ヤオコーの税込価格は小数点以下 2 桁まであるため，税込価格は右から 2～3 桁目の間隔X_{23}が広い．一方，税抜価格は“, ”が入るため右から 3～4 桁目の間隔X_{34}が広くなる．両者の関係を調べたところ，税込価格では$X_{23} > X_{34} + 1$が成り立つことが分かったため，この条件を用いて税込／税抜を識別した．

5. 実験

ヤオコー公式 HP からダウンロードしたチラシ画像（42.3cm×59.9cm，240dpi，4000×5659 画素）を 6 面分用い実験を行った．パラメータは実験により最適化し，全ての画像で同じ値を用いた．1071 個ある価格領域のうち円領域が正しく認識されたのは 1067 個，更に価格の数値が正しく認識されたのは 1064 個であった．価格であると誤認識された非価格領域の数は 5 個であり，適合率 99.53%，再現率 99.35%，F値 99.44%が得られた．また，税込／税抜の識別正解率は 100%であった．認識失敗の原因は，「漢字の一部を数字と誤認識」，「傾いている数字を誤認識」などである．

今後はこれら誤認識を改善するとともに，商品名の認識，及び価格と合わせて商品情報のデータベースを生成する機能を実現したい．



(a)原画像　　(b)認識された価格領域

図 1　実験結果の一部

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/.