修飾文字の背景除去に関する一検討

6R－6

久保田 哲也　糸井 清晃　新井 浩志　小林 幸雄
千葉工業大学 電子工学科

１．はじめに

新聞などの文字情報は、コンピュータに文字認識させ文字符号にして保存することにより、情報量を少なくすることが出来る。また、文字符号にすることにより情報検索も容易に行えるようになる。しかし、文書内容の検索に必要な新聞の見出し部分には、文字の背景に飾りが含まれていることがあるために、このことが文字認識を困難にしている。そこで、修飾文字の背景を除去し、文字部分のみを抽出する方法を報告する。

２．対象（新聞の見出し）画像の検討

スキャナーで読み込んだ画像１００種類に対して、フィルター処理及び２値化を行い、分類した。

a.２値化で文字が抽出できる見出し(70%)
b.背景にグラデーションがある見出し(10%)
c.文字の片側に飾りがある見出し(15%)
d.その他(5%)

３．修飾文字の背景除去

上記のように分類した見出しについて、それぞれ修飾文字の背景除去法を検討した。その他の見出しに関しては、検討を行わなかった。

３－１．２値化で文字が抽出できる見出し

２値化を行うことで文字の抽出は可能である。したがって、２値化する際の閾値を自動的に決定する方法として、判別分析法について検討した。

本研究では、一度判別分析法を用いて閾値ｔを求め、次に最小濃度値からｔまでの濃度、ｔから最大濃度値までの濃度で再度判別分析法を行い、全体で３つの閾値を求め２値化した。

Background Pattern Elmination of Decorative Character.

Tetsuya Kubota,Kiyoaki Itoi,Hiroshi Arai, Yukio Kobayashi

Department of Electronics, Chiba Institute of Technology

３－２．背景にグラデーションがある見出し

以下に示す処理を行った。

1.判別分析法を用いて閾値を決め、２値化する。
2.細かい雑音を除去する。
3.白画素に対してラベリングを行う。
4.一番外側の画素から３画素以内を検索し、ラベリングされた領域が見つかった場合、その領域を背景とし黒く塗りつぶす。

３－３．文字の片側に飾りがある見出し

文字と文字・背景と背景は隣り合わないという特徴を利用して、次の処理を行った。

1.判別分析法を用い閾値を決め、２値化する。
2.白画素に対してラベリングを行う。
3.それぞれのラベルの周りを調べ、文字であるのか背景であるのか判断する。
4.背景と判断されたラベル領域を、黒く塗りつぶす。

［ラベリング後の検索方法］

図１を参照に説明する。必ず①には背景、②には文字がくる。そこで、③以降に対して検索を開始する。背景が黒い半分の領域では、ラベリングされた領域は全て文字である。したがって、この領域に存在するラベル名を②に書き換える。次にラベリングと同様の方向で検索をし、最初に①・②以外のラベル領域を見つけたところで、周りの画素の検索を行う。その点から上下左右４方向を調べると、周りには文字または背景が存在する。４方向のうち２カ所以上背景を見つけたら、その領域を文字と判断し、ラベル名を②に書き換える。また、背景が２カ所以上見つからなければ背景と判断し、ラベル名を②に書き換える。その他のラベルに対しても同様に行うことにより、ラベル名が①と②だけとなり、文字（②）と背景（①）に分類することが出来る。

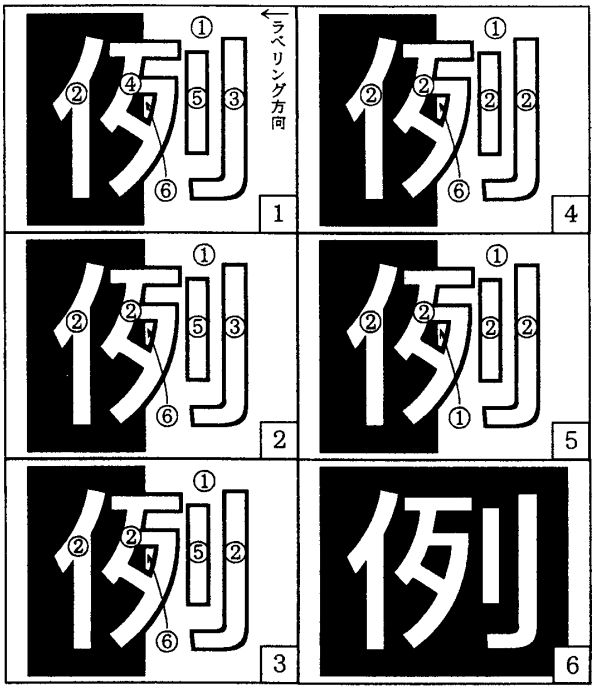

図１　文字抽出例

４．実験結果及び考察

２値化で文字が抽出できる見出し１００種類を対象に、判別分析法を用いて閾値決めの自動化を行った。本研究では３つの閾値を求めたが、この閾値のいずれかで１００種類全ての見出しの文字を抽出することが出来た。しかし、３つの閾値からの最適な閾値の選択は手動で行うため、完全自動化とすることは出来なかった。

背景にグラデーションがある見出し５０種類を対象に、３－２．で述べた方法を用いて実験を行った。完全に文字が抽出できた物（図２-(a)）が７４％、若干文字が欠けてしまう物が１０％、大きな雑音が残ってしまう物（図２-(b)）が１６％（新聞全体でから見ると、約２％程）という結果を得た。若干文字が欠けてしまう物も含めると、８４％の見出しの文字が抽出できたことになる。

文字の片側に飾りがある見出し５０種類を対象に、３－３．で述べた処理を行った。その結果、完全に文字が抽出できた物（図２-(c)）が９０％、若干文字が欠けてしまう物（図２-(d)）が１０％となった。若干文字が欠けてしまう物は、文字の点が一つか二つ消えてしまう程度であるので、文字認識できるものと思われる。よって、文字の片側に飾りがある見出しに関しては、この処理方法によってほぼ１００％の修飾文字を抽出できたことになる。新聞全体としては、本研究で行わなかったその他の見出しを含めても、９０％程の修飾文字の検出ができた。

今後の課題としては、実際に抽出できた文字を文字認識装置にかけてみてどの程度認識できるのかを確認する必要がある。また、その他の見出しと、グラデーションの見出しに関しての再検討が必要である。そして最終的には、全ての処理を自動的に行えるようにする必要がある。

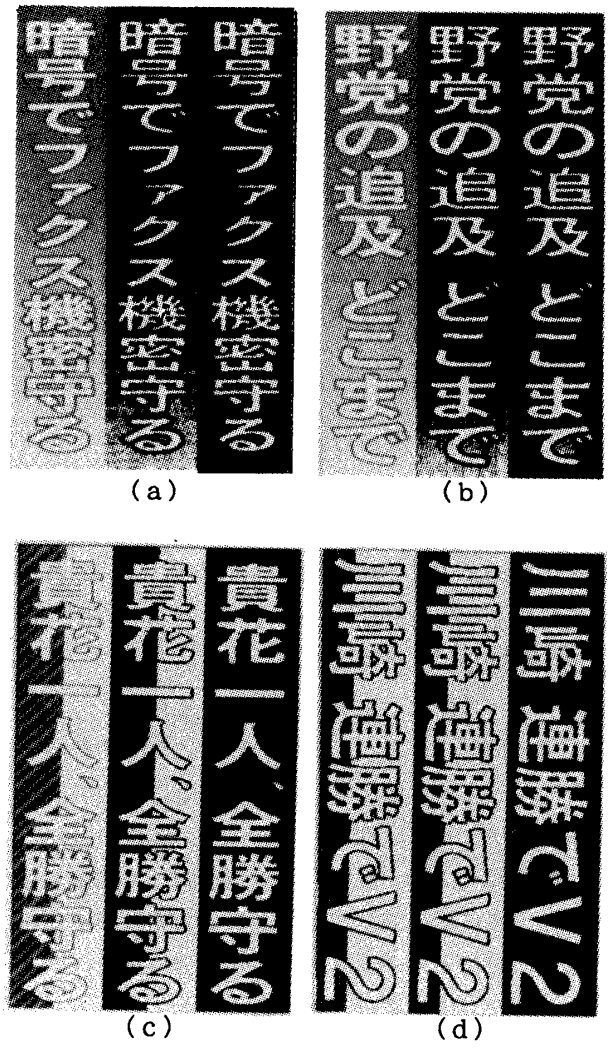

図２　実行結果

５．参考文献

１）画像処理とパターン計測技術，日本機械学会編，朝倉書店

２）林，高井，成田：画像処理による飾り文字の復元，電子情報通信学会論文誌，PRU94-12