岡山理科大学紀要 第38号 A pp. 95-103(2002)

カラー文書と濃淡文書からの文字列抽出の比較実験

幸 弘道・横山義生*・鷲田 智**・濱本高志***・大倉 充***

岡山理科大学大学院工学研究科情報工学専攻
*神戸市立平野中学校
**岡山情報処理センター
***岡山理科大学工学部情報工学科

(2002年11月1日 受理)

1. まえがき

文書画像中の文字を認識する場合，その初期段階における文字列抽出処理はその後の処理に大きく影響を与えるため重要な処理の一つと位置付けられる．そのため，これまでに多くの研究が行われてきた[1]－[6]．カラー文書と濃淡文書からの文字列抽出を考えた場合，当然，明度情報しかない濃淡文書に比べ，色の情報があるカラー文書からの文字列抽出の方が容易だと考えられる．しかし，これまでの研究では，カラー文書と濃淡文書では個別の文字列抽出手法の提案が多く，同一の文書をカラーおよび濃淡画像としてコンピュータに入力して文字列抽出の比較を行った例は少ない．

本研究では，どの程度の抽出率の差が生じるのかを調査するために，雑誌の表紙等の複雑な背景を有する文書をカラー画像および濃淡画像としてコンピュータに入力し，各々に対して文字列抽出処理を施し抽出率の比較検討を行った．これら2種類の画像に対して，文字列候補領域の抽出にはそれぞれ異なる処理を，文字列の抽出には同一の処理を行った．

2. 実験概要

本研究では，全ての作業をパーソナルコンピュータを用いて行った．雑誌広告等をイメージスキャナ（ES-8000：EPSON）を用い，画像サイズに応じて解像度を 72dpi から 150dpi として，RGB 各 256 階調のフルカラー画像としてコンピュータへ入力する．入力された画像に対する文字列候補領域及び文字列領域の抽出処理はC言語で記述した．実験に用いた画像は濃淡文書およびカラー文書各 16 枚である．その一例を図1に示す．

(a) (b) (c)

図1 実験に使用したサンプル画像

3. 処理の概要

3.1 文字列候補領域の抽出

本研究では，カラー画像の場合には均一色からなる領域，濃淡画像の場合には均一濃度からなる領域ごとに画像を分解する処理を文字列候補領域の抽出処理と位置付ける．

3.1.1 カラー画像に対する文字列候補領域の抽出処理

カラー画像に対しては，まず代表色（画像中に多く存在する“主要な色”）を以下の手順で選出する．RGB色空間を各軸均等分割して小領域を作成した後に，画像の各画素がどの小領域に属するかという画素ヒストグラムを求める．画素頻度が極大値を示す小領域を検出して，その小領域が示す色を初期の代表色とする[1]．次に得られた代表色を基にRGB色空間でクラスタリングを行い画像全体を数種類の代表色により減色する．クラスタリングにはK－平均法を用いた[7]．K－平均法により，初期の代表色をクラスタ中心として，これと全ての画素に対しユークリッド距離を用いて距離を求め，最も距離が近い画素から構成されるクラスタを求める．本研究におけるユークリッド距離は式(1)に示すように画像の1画素Gの有するRGB値と，K個の代表色（C_1～C_k）の有するRGB値間で計算を行う．

画像中の1画素： $G(R_0,G_0,B_0)$

K個の代表色： $C_k(R_k,G_k,B_k)$

$$D_k=\sqrt{(R_k-R_0)^2+(G_k-G_0)^2-(B_k-B_0)^2} \tag{1}$$

次に各クラスタに所属する画素の平均値から新たなクラスタ中心を生成する．この処理を現在のクラスタ中心と前に得られたクラスタ中心とが変わらなくなるまで繰り返すことによって，最終的に得られたクラスタ中心をその画像の代表色とする．最後に各代表色ごとに2値画像を生成する．生成された2値画像の例を図2に示す．

3.1.2 濃淡画像に対する文字列候補領域の抽出処理

カラー画像からの濃淡画像の生成はAdobe Photoshop6.0を用いた．カラー画像から明度情報やG（緑）成分のみを抽出して濃淡画像を生成しいくつか処理を行ったが，Adobe Photoshop6.0を用いた場合の結果が最も良好で安定した結果を示したためである．

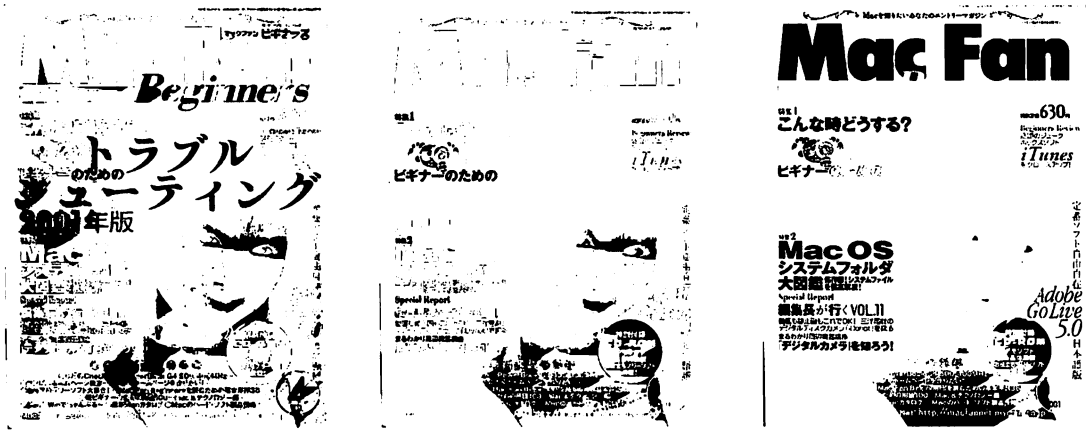
濃淡画像に対しては，まず，一つの領域内の濃度の変動を抑えるために領域併合法を行う[8]．領域併合法により，画像を小さな領域に分け，一様と見なされた領域の濃度平均値を求め新たな領域を作成する．この処理を領域の変更がなくなるまで繰り返し，一つの領域の画素濃度を均一にする．領域の濃度が均一化された画像の濃度値を8段階に再量子化し，濃度値別に8枚の2値画像を得る．ただし，領域が2つの画像に分断されることを防ぐために，濃度値に重なりを持たせて画像の分割を行う．分割する際の濃度値の範囲を表1に示す．生成された2値画像の例を図3に示す．濃度値に重なりを持たせたために重複して抽出される領域が数多く存在する．

表1 濃淡画像における2値画像生成時の濃度値の範囲

画像	1	2	3	4	5	6	7	8
濃度値	0～47	16～79	48～111	80～143	112～175	144～207	176～239	208～255

(a) (b) (c)

図 2 文字列候補領域の抽出処理により生成された 2 値画像の例（カラー画像）



(a) (b) (c)

図 3 文字列候補領域の抽出処理により生成された 2 値画像の例（濃淡画像）

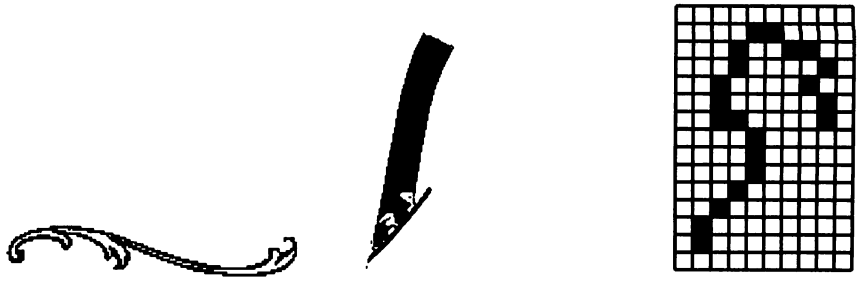
3.2 文字列抽出処理

本研究では，3.1 節で得られた複数枚の 2 値画像それぞれに対して文字列抽出処理を行う．本研究では，文書画像に含まれる文字は以下のような条件を持つと仮定した．

- ある一定サイズ内の大きさを持つ．
- ほぼ矩形で囲まれる．
- 文字が単独に存在する事は稀である．
- 1 つの文字列は文字同士が隣接し，互いに同じ大きさで書かれている．

上記した条件を考慮に入れ，ラベリング処理[8]を施し，それぞれの領域で雑音と思われる部分を除去する．図 4 に雑音と判定した領域の例を示す．同図(a)は領域にホール（孔）が多数存在する例，(b)は非常に細い線分，(c)は領域の縦横比に極端な差がある例である．この他に画像に点在する孤立点や画像サイズと比較してその領域が文字と考えられない程大きいものは文字ではないとして除去する．全ての処理後，分割していた 2 値画像を統合し，残った領域を文字列抽出結果とする．図 5 に文字列抽出結果の一例を示す．

(a) 領域にホール（孔）が多数存在する例



(b) 領域の縦横比に極端な差がある例　　(c) 微細な線分

図 4　文字ではないと判定した領域

(a)原画像　　(b)濃淡文書画像　　(c)カラー文書画像

図 5　文字列抽出結果の一例

4. 実験結果

4.1　文字列候補領域の抽出実験結果

文字列候補領域の抽出において，カラー画像と濃淡画像それぞれに対する結果を表 2 に示す．表には図 1 に示した画像(a)，(b)，(c)に対する結果と使用した 16 枚に対する合計の結果を示している．文字列の分裂とは図 6 に示すように一つの文字領域が 2 値画像に分割した際に分かれて出力されることであり，文字と背景の融合は図 7 に示すように背景領域と文字領域が一つの領域になってしまうことである．分割した画像において，文字が欠けていたり，背景領域と融合してしまったものは，すでにこの時点で抽出に失敗である．

濃淡画像においては，濃度値に重なりを持たせて 2 値画像を生成しているために，文字領域と背景領域との融合が多く見られた。分裂した文字列の数は，領域併合法により，文字領域と背景領域の境界部分の濃度が均一化してしまい，文字の輪郭部分の濃度が文字自体の濃度と異なる値を示したために，細い文字が欠けてしまうことで多くなったと考えられる．また抽出失敗と判定してはいないが，文字の一部が欠けていたり，逆に潰れていたりと，抽出処理や認識処理において，良好な結果を得る事が難しいと考えられるものがカラー画像に比べ多く存在した．カラー画像においては，代表色が多数現れるときに淡い色の小さな文字が図 6 のように分裂してしまう部分が存在した．

表 2　文字列候補領域の抽出に失敗した文字数（画像数 16）

画像	濃淡文書		カラー文書	
	文字列の分裂	文字と背景の融合	文字列の分裂	文字と背景の融合
(a)	26	47	0	14
(b)	1	0	0	0
(c)	62	0	58	0
合計 16 枚	241	81	94	23

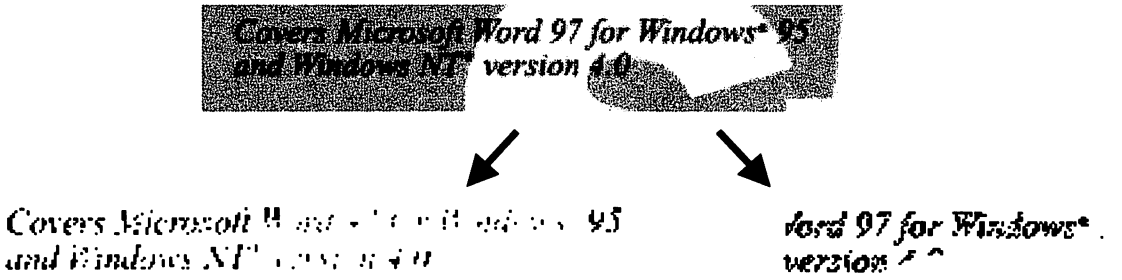


図 6　文字列の分裂

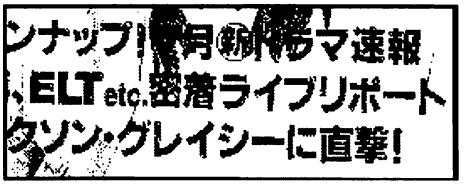


図 7　背景と文字との融合

4.2　文字列領域の抽出実験結果

文字列抽出処理の結果を表 3 に示す．表中の括弧内の数字は画素数を示す．80×80 画素以上の大きな文字に関しては，ほぼ全ての画像においてかなり良好な抽出結果を示している．抽出できなかった文字は図 8 に示す文字同士が連結しているものや図 9 に示すような文字領域に他領域が重なっている部分のみであった．最も大きな差が生じたのは 30×30 画素以下の小さな文字で，濃淡画像の抽出結果はカラー画像より 15%も悪かった．さらに，抽出できたと判定した文字に関しても，カラー画像に比べて濃淡画像ではつぶれ等の文字の劣化がより多く見られた．

最終的に残った雑音数を表 4 に示す．表からも分かる通り，雑音数は全てカラー画像より濃淡画像の方が多くなった．特に濃淡画像においては 2 値画像を統合したときに，文字が抽出できているにも関わらず背景領域と重なってしまい，読めなくなっている部分が存在した．カラー画像と濃淡画像の文字抽出率の違いはこの除去できなかった雑音が大きな要因の一つと考えられる．

表 3　文字列抽出結果（文字数）

画像	原画像			濃淡文書			カラー文書		
	大(80×80)	中(50×50)	小(30×30)	大	中	小	大	中	小
(a)	26	132	370	9	111	235	18	120	324
(b)	36	194	19	23	160	6	36	186	19
(c)	19	102	81	16	46	8	17	81	21
合計 16 枚	410	1634	2157	344	1372	1449	387	1509	1829

表 4 除去できなかった雑音領域の数

画像	濃淡文書			カラー文書		
	大(50×50)	中(30×30)	小(10×10)	大(50×50)	中(30×30)	小(10×10)
(a)	0	7	6	0	3	2
(b)	0	3	14	0	2	3
(c)	1	1	1	0	3	1
合計 16 枚	31	65	107	12	52	58

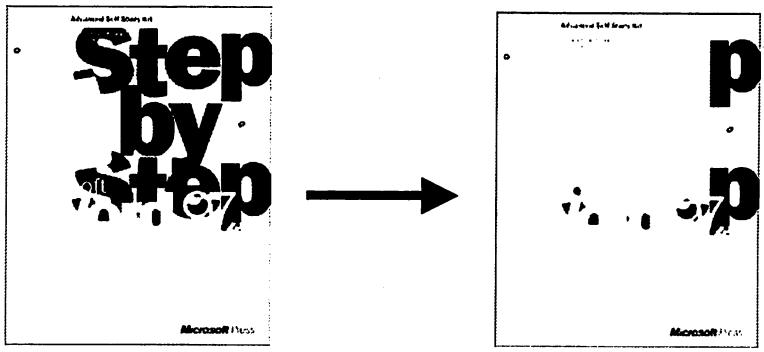


図 8 文字同士が連結し，非常に大きな領域となった例

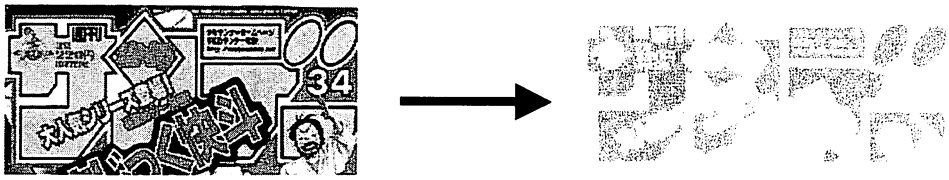


図 9 文字領域に他領域が重なっている例

5. 考察

全体の文字抽出率の結果を表 5 に示す．実験前に予想したように，明らかにカラー画像に対する結果が濃淡画像に対する結果よりも良好であった．濃淡画像よりカラー画像の抽出率は 10%程度高く，残った雑音も約 50%であった．個別に画像を調査しても，濃淡画像がカラー画像よりも良好な結果を示すことはまずなかった．ほぼ等しい抽出率を示した場合においても，雑音がカラー画像に対するものよりも多く残っていた．

50×50 画素以上の大きな文字に関しては，カラー画像では 90%以上の抽出率であったが，濃淡画像では 80%程度の抽出率であった．30×30 画素以下の小さな文字に関しては，カラー画像で 80%程度，濃淡画像では 65%程度と大きな文字に比べて差が大きかった．なお，抽出できたと判定した文字に関しても，カラー画像に比べて濃淡画像ではつぶれ等の文字の劣化がより多く見られた．

表 5 文字列抽出率（%）

画像	濃淡文書				カラー文書			
	大	中	小	全体の抽出率	大	中	小	全体の抽出率
(a)	35	84	64	67	70	91	88	88
(b)	64	83	32	75	100	96	100	97
(c)	84	45	1	35	89	79	26	59
平均	84	84	67	75	94	92	83	88

5.1 抽出に失敗した例

図 10 は濃淡画像からの文字列候補領域の抽出に失敗した例である．文字領域と背景領域がほぼ同じ濃度差を示している．元のカラー画像では背景色と文字色は明確な違いを示していたが，濃淡画像に変換すると視覚的にも判別しづらくなってしまう．図 11 はカラー画像からの文字列抽出に失敗した例である．文字領域の色と背景領域の色が近いために，カラー画像において背景領域と文字領域が融合してしまう．カラー画像において文字領域が融合する部分は濃淡画像においても融合する場合が多い．図 10，11 に示した例のような領域が多数存在すると文字抽出率は極端に低下することになる．

図 12 は，文字の一部分が細いために候補領域抽出で文字が欠けてしまった例である．これらの領域は抽出失敗と判定してはいないが，文字認識の段階で誤認識を示す可能性が高く，また文字列抽出においても雑音と判断され除去される可能性がある．

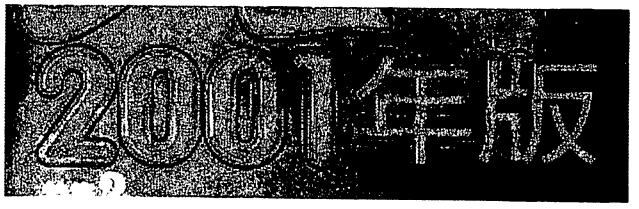


図 10 背景領域と文字領域の濃度がほぼ同じ例

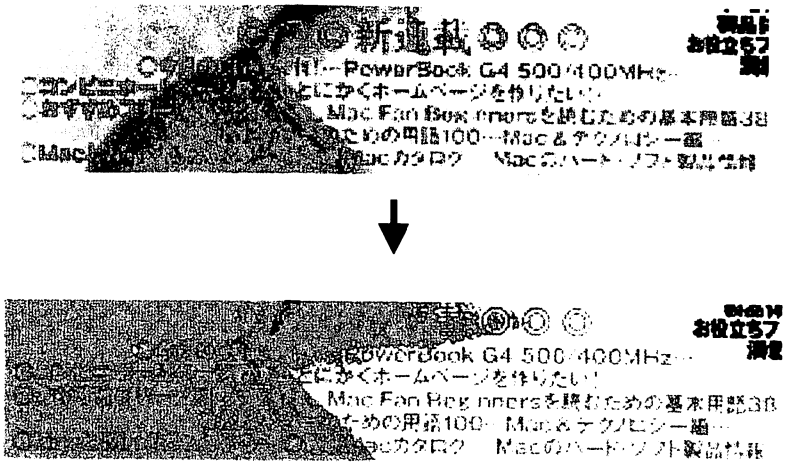


図 11 背景領域と文字領域の色が近い例

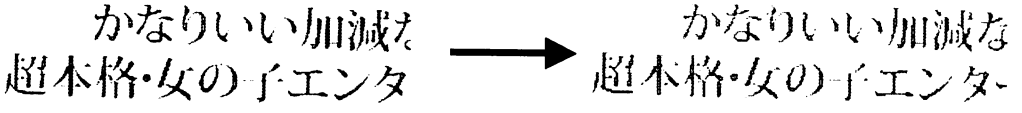


図 12 輪郭が細いために文字の一部分が欠けた例

5.2 処理時間

図 2，3 の 2 値画像が示すように，濃淡画像とカラー画像では候補領域抽出処理から得られる 2 値画像に大きな違いがある．カラー画像においては背景領域と文字領域の色は区別されているために，2 値画像においてそれらの領域は明確に区別されるが，濃淡画像においては一枚の画像に背景領域と文字領域が混在し，画像に占める領域部分が非常に多くなる．これにより表 6 に示すように，濃淡画像の方がカラー画像と比べ，雑音除去の処理時間が長くなる．また画像に占める領域が多いということは，同じ雑音除去処理を用いていることから，カラー画像よりも濃淡画像の方が残る雑音の量が明らかに多くなってしまう．

表6 平均処理時間（秒）

濃淡文書		カラー文書	
文字列候補領域抽出処理	文字列抽出処理	文字列候補領域抽出処理	文字列抽出処理
176.9	257.1	12.8	146.7

6. おわりに

本研究では，カラー文書と濃淡文書からの文字列抽出に関する比較実験を行った．カラー文書では濃淡文書に比べ約10%高い抽出率が得られ，雑音も濃淡文書の約50%という結果が得られた．濃淡文書では，文字列と背景領域との濃度差が小さい場合，現状では明確な分離が困難なために濃度値に重なりに持たせて候補領域を求めるが，これが雑音領域の数を増加させてカラー文書との結果の差を著しくしている．カラー文書も濃淡文書も抽出率においては良好な結果が得られたが，まだ雑音領域が多く残ってしまう．今後は，文字領域と雑音領域の区別処理の検討と文字列候補領域抽出処理の高精度化に取り組み，カラー文書と濃淡文書の両者に適用可能な文字列抽出法を考案したい．

参考文献

1) 篠川敏行，長谷博行，米田政明，坂井充，丸山博，“カラー画像中の文字列抽出”，信学技報，PRMU96-71，1996.
2) 上羽葵，武田哲也，岡田至弘，“等色線処理によるカラー画像からの文字列領域の抽出”，画電学誌，vol25，no4，pp.382-389，1996.
3) 仙田修司，美濃導彦，池田克夫，“文字列の単色性に着目したカラー画像からの文字パタン抽出法”，信学技報，PRU94-29，1994.
4) 後藤英昭，阿曽弘具，“文字行の局所的な直線性を利用した頑健・高速な文字行抽出法”，信学(D-II)，vol.J.78-D-II，no.3，pp465-473，March 1995
5) 後藤英昭，平山理継，阿曽弘具，“局所多値しきい値処理法による濃淡文書画像からの文字パターン抽出”，信学(D-II)，vol.J.82-D-II，no.11，pp2188-2192，Nov 1999
6) 長谷博行，米田政明，坂井充，丸山博，“カラー文書画像中の文字領域抽出を目的とした色分割についての検討”，信学(D-II)，Vol.J.83-D-II,no.5，pp.1294-1304，May 2000.
7) 手塚慶一，北橋忠宏，小川秀夫，ディジタル画像処理工学，オーム社，1992.
8) 長谷川純一，輿水大和，中山晶，横井茂樹，画像処理の基本技法，技術評論社，1990.

Comparison of Character String Extraction from a Color Document and Shade Document

Hiromichi YUKI, Yoshio YOKOYAMA*, Satoshi WASHIDA**,

Takashi HAMAMOTO*** and Mitsuru OHKURA***

Graduate School of Engineering,
Okayama University of Science,
1-1 Ridai-cho , Okayama 700-0005, Japan
**HIRANO Junior High School.*
***Okayama Electronic Data Processing System Center Co., Ltd.*
****Department of Information & Computer Engineering,*
Okayama University of Science

(Received November 1, 2002)

When the extraction of character strings from color documents and shade documents is considered, color documents are usually easier to process. This is because a shade document has only brightness information. In previous research, color and shade documents have been processed using different methods. However, there is little research which has compared the extraction rate of a single method for both types of document. In this research, we investigate the difference in the rate of extraction using a novel processing method. We first generated a color and shade image and input these into a computer. The document used for generating these images had a complicated background. Next, we performed character string processing on the images and compared the rates of extraction. The result of the experiment showed the extraction rate was 10% higher for the color image than for the shade image.