

モルフォロジによる複雑レイアウト画像からの 文字列抽出に関する研究

田 中 直 樹 神戸商船大学商船学部助教授

1. はじめに

雑誌や書籍の表紙には書名、著者名、巻号および出版社名などが記されている。これらは、書誌情報と呼ばれその書籍の内容に関する重要な情報を担う。従って、書籍の電子データ化の際にはこれらの情報を抽出することが不可欠となる。雑誌などの本文ページは文字や図表などで構成されるシンプルなレイアウトを持ち、これらのページを観測した画像は「文書画像」と呼ばれ、それらの画像の文字列、図表、背景部の分離は比較的容易である。これに対して、雑誌等の表紙は、背景に風景や人物の写真、イラストを配されるなどし、またそれらと文字列、あるいは文字列同士で重なり合うなどかなり複雑なレイアウトであることが多い。そのために、文書画像の場合と比べて背景部と文字列の分離処理がこれまで極めて困難とされてきた。顧らは、モルフォロジを用いた文字領域抽出方式を新たに提案している。提案方式は、まず差分 Top-hat 演算により、領域幅に基づくセグメンテーションを行う。さらにモルフォロジ演算などにより、文字の骨格抽出、領域成長などにより文字領域抽出を行う。雑誌表紙画像や文字列を持つ情景画像を用いた文字領域抽出実験を通じて提案方式がそれら画像からの文字領域抽出に有効であることを示している。しかしながら同方式では、各抽出回数における2値化、抽出回数間の文字列の競合解消などが未解決の問題として残されていた。本稿では、これらの未解決部分の研究を進め、部分領域2値化と領域競合解消方式を新たに導入した文字列抽出方式について報告する。

2. 文字列抽出方式

提案文字列抽出方式は、抽出される文字列の文字サイズに応じて2つの処理に分かれる。文字サイズの小さな文字列に対応するものを文字列抽出方式1、文字サイズの大きな文字列に対応するものを文字列抽出方式2と呼ぶ。抽出方式1、2の差異は、主に方式2がCannyエッジ検出器を用いている点である。また、従来の方式では横方向の文字列のみが抽出対象であったが、本提案方式では、文字の方向を限定することなくどの方向の文字列も抽出可能となっている。

2.1 文字列抽出方式1：文字サイズ小

2.1.1 領域分割

これまで顧らにより式(1)・式(2)のような方式が開発されてきた。

文字の幅が21画素以下の文字列を抽出するための第一段階である領域分割処理は次式により行う。

$$\left\{ \begin{array}{l} X_i = |(X_0 - X_0 \circ r_i \text{Bdisk}) - (X_0 - X_0 \circ r_i \text{Bdisk})|_B - X'_{i-1} \\ X'_i = \bigcup_{0 < j \leq i} X_j, X'_0 = \phi \end{array} \right. \quad (1)$$

$$\left\{ \begin{array}{l} X_i = |(X_0 - X_0 \circ r_i \text{Bdisk} - X_0) - (X_0 \circ r_{i-1} \text{Bdisk} - X_0)|_B - X'_{i-1} \\ X'_i = \bigcup_{0 < j \leq i} X_j, X'_0 = \phi \end{array} \right. \quad (2)$$

ここで、

$i = 1, 2, 3, \dots, 10$

X_0 ：濃淡画像の原画像

$| \cdot |_B$ ：特定のしきい値（濃淡画像では $B=2$ ）で2値化処理

式(1)・式(2)を入力画像に対してそれぞれ独立に適用する。式(1)の X_1 には riBdisk の直径に等しい幅をもつ、周囲より浮き出した(大きい階調値=白い)領域が順次求まる。式(2)も同様であるが、ここでは周囲よりも沈み込んだ(小さい階調値=黒い)領域が求まることになる。処理手順については、半径1の円盤型構造化要素から初めて、一連の画像 X'_i を漸化的に生成する。処理は、処理対象画像中に含まれる文字のうち21画素以下の幅をもつ領域が抽出される。



Fig.1 X_0 : 原画像



Fig.2 X'_5

2.1.2 主要文字線の抽出

2.1.1 で得られた領域画像に文字領域とノイズが共存している。よってノイズを除去した，文字線領域の主要部分を抽出するために，次式による処理を行う。

$$\begin{aligned} E'_i &= (((X_i \circ r_{i-1} \text{Bdisk}) \cdot r_{i-1} \text{Bdisk}) \circ r_{i+2} \text{Bdisk}) \oplus r_i \text{Bdisk} \\ E_i &= E'_i \cap X_0 \end{aligned} \quad (3)$$

E_i の例を Fig.3 に示す。

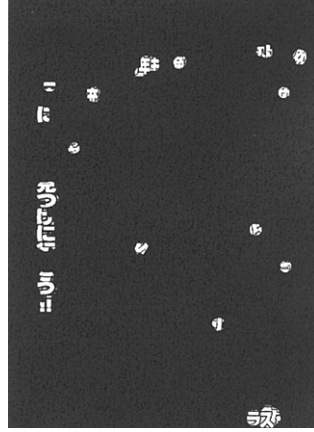


Fig.3 E_5

2.1.3 文字列候補領域抽出

上式で生成されたマスク領域の反 E'_i について，Dilation を施して文字領域を拡大した後 Closing 処理により穴埋めを行う。さらに背景部に対して微小領域除去を行う。(Fig.4)

$$\begin{aligned} F'_i &= T_{1000} [(\overline{E'_i \oplus r_{6+i} \text{Bdisk}}) \cdot r_{1+i} \text{Bdisk}] \\ F_i &= \overline{F'_i} \end{aligned} \quad (4)$$

Tn : n 画素以下の領域を削除 (n は自然数)

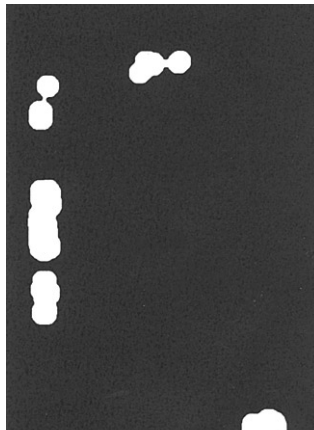


Fig.4 F_5

2.1.4 小領域除去

文字列候補領域には，多くの場合，文字列でない領域が存在している。経験的に，領域が小さく，付近に別

の文字列候補領域がない場合は、ノイズである可能性が高い。そういった領域に対し、一定画素数以下の領域についてはノイズとして除去する。

$$D_i = T_{1400 + (i \times 300)} |F_i \cdot r_{i+1} \text{Bdisk}| \quad (5)$$

2.1.5 競合の解消

各次数で抽出された文字列領域 (D_i) は、複数の次数に同じ文字列が含まれている場合がある。その競合を解消するために、次のアルゴリズムを用いる。まず有効な各次数の領域に対してそれぞれの位置関係を求め、競合している部分があるものに対しては、その競合している部分の面積を求め、両領域との割合を出す。その割合のいずれかがしきい値 (75%) 以上であり、かつ両領域の次数の差が 2 以内であれば、割合の大きい方の領域を削除し、2 以上なら割合の小さい方の領域から大きい方の領域を削除する。(Fig.5)

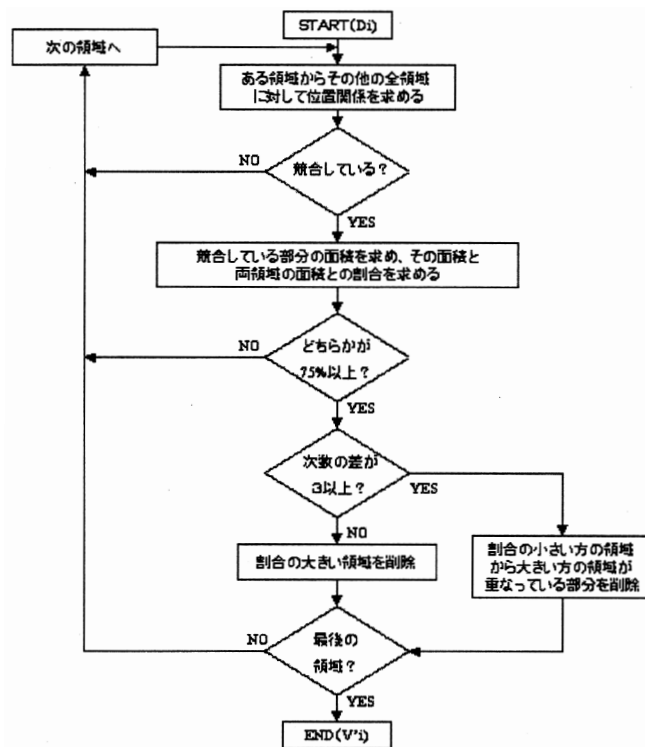


Fig.5

2.1.6 部分ヒストグラム法による 2 値化

2.1.5 の競合解消により得られた各次数の 2 値画像の集合を V_1 とし、原画像との積演算 (マスク処理) によって得られる濃淡画像の集合を V_1 とすると、

$$V_i = v_i \cap X_0 \quad (6)$$

と表すことができる。次に E_i に対してそれぞれの文字列候補領域についてヒストグラムを求め、各領域のピーク (P_i) と P_i に隣接する階調値の一つを採用する。(隣接する階調値の採用にあたっては、画素数の多い方を採用) これらの値を V_i のそれぞれ対応する部分に適用し、同じ階調値をとる画素を文字列領域 (L_i : 階調値 0 と階調値 i をもつ 2 値画像) として検出する。(Fig.6)

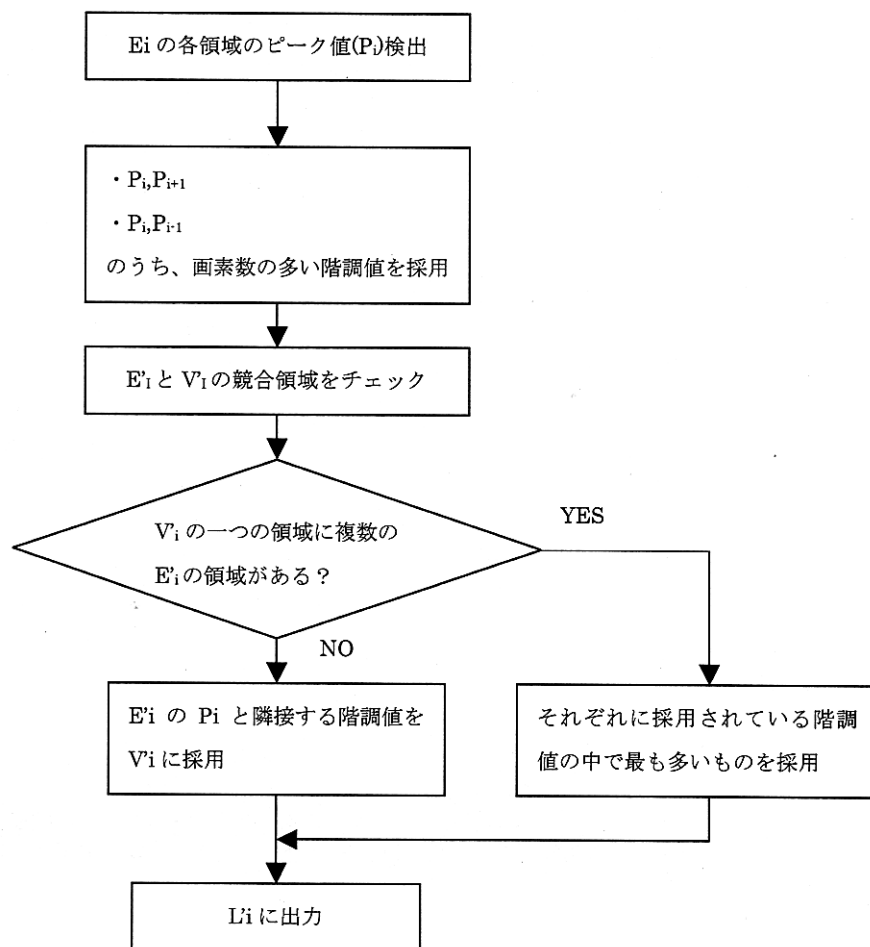


Fig.6

2.1.7 抽出文字列の整形

抽出された L'_i に対して、次に示す微小領域除去式(7)と Conditional Dilation (式(8), 式(9)) を用いてノイズ除去処理を施す。

$$K_i = T_3[L'_i] \quad i \leq 2 \quad (7)$$

$$\begin{cases} K_{i0} = L'_i \circ r_{i-2} \text{Bdisk} \end{cases} \quad (8)$$

$$\begin{cases} K_i = K_{i0} \oplus |L'_i| \text{Bsquare} \quad 2 < i \leq 10 \end{cases} \quad (9)$$

最後の処理で得られた領域画像 $K_i (i = 10)$ が文字列抽出結果画像となる。表示のために各次数の総和を一枚の画像にしたものが Fig.7 である。

$$Z'_{r1} = \bigcup_{1 \leq i \leq 10} K_i \quad (10)$$

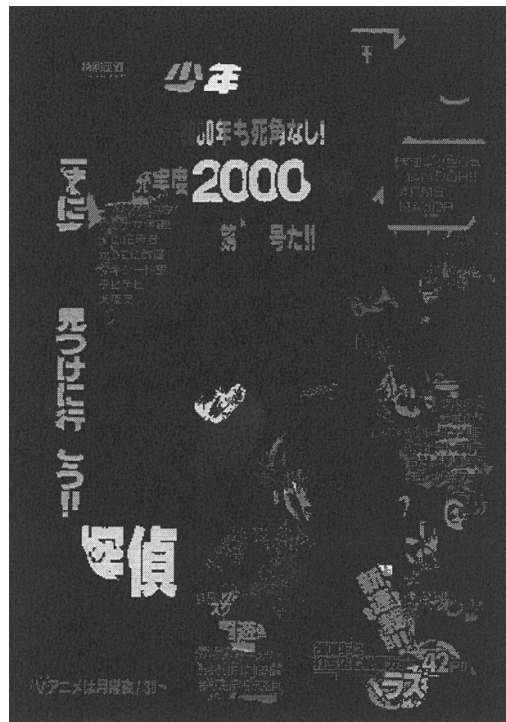


Fig.7 Z'_{r1}

2.2 文字列抽出方式2：文字サイズ大

2.2.1 Canny エッジ検出

ここでは、まず Canny エッジ検出器を用いて原画像 X_0 からエッジ Y を検出する。今回 Canny エッジ検出器を使用するにあたって、ユーザパラメータを以下のように設定し、エッジを検出した。(Fig.8)



Fig.8 Y

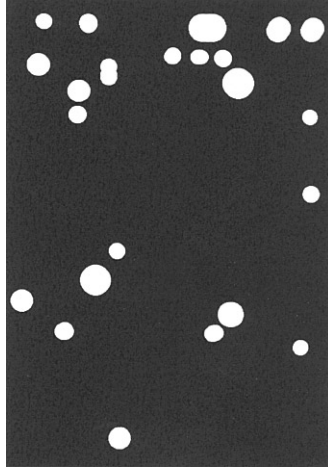


Fig.9 S'_{17}

$\sigma : 1.0$

high threshold : 0.5 (しきい値の上限)

low threshold : 0.3 (しきい値の下限)

2.2.2 文字列候補領域抽出

Canny エッジに対して、式(11)、式(12)の処理を施す。この処理により、ある範囲の領域幅を持つ領域の主要な部分が抽出される。なお本実験での入力画像で、文字線の最大幅は70画素を越えないと仮定し、 i の値は最大34までとした。また $i=1,2,3\cdots 10$ は、2.1の処理ですでに文字列を抽出しているため、ここでは処理の対象とならない。

$$S'_i = ((Y \circ r_{i-1} \text{Bdisk}) - (Y \circ r_i \text{Bdisk})) \cdot r_{i-1} \text{Bdisk} \quad (11)$$

$$S_i = \bigcup_{i \leq j \leq 34} S'_j \quad (12)$$

$i=11,12,13\cdots 34$

Y : Canny エッジ画像

得られた S'_i は特徴を強調した2値画像である。

(Fig.9 は S'_{16} から S'_{34} までの論理和である。)

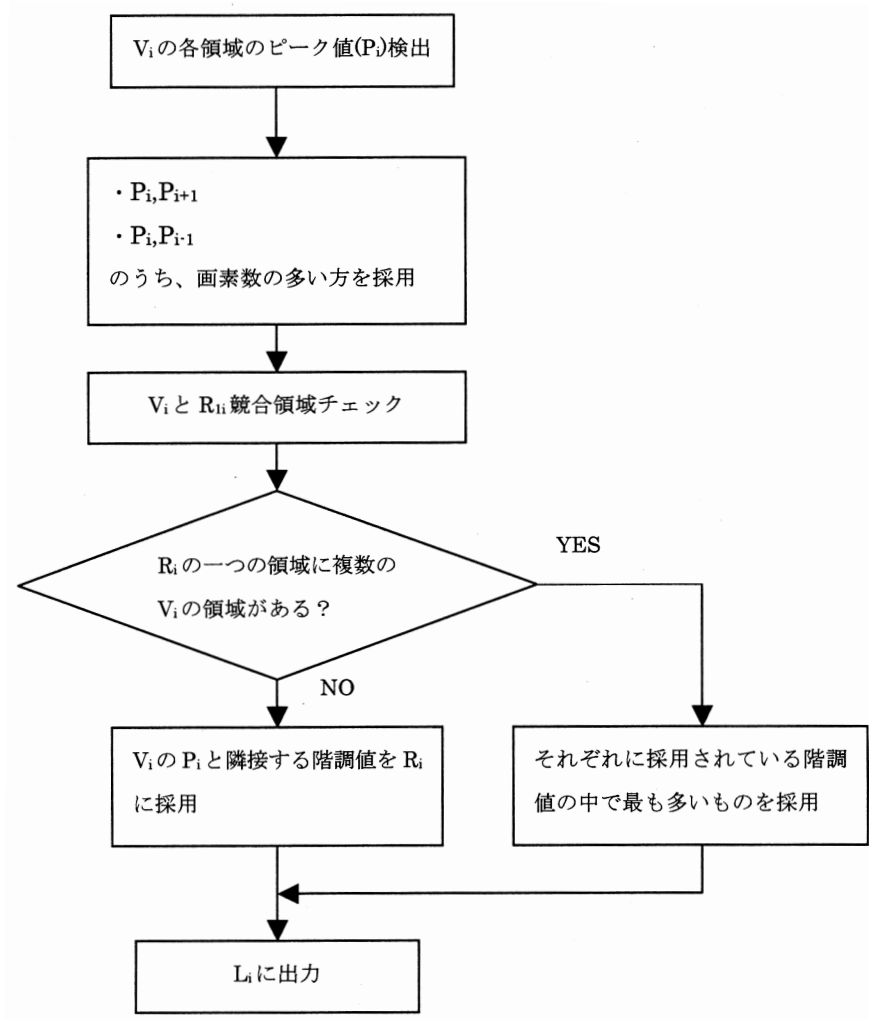


Fig.10

2.2.3 文字列領域抽出

2.2.2 で得られた S_i は S'_i の論理和であり、各次数で検出されたものが累積されるため、全ての S_i から文字列を抽出する必要はない。したがっていくつかの S_i を選択し以下の文字列領域抽出処理を施せばよい。予備的な実験により、空でない S_i の最大次数から2次数置きに採用すれば十分であることが分かった。例示している画像では $i=17,14,11$ となる。

$$V_{17} = S_{17} \cap X_0 \quad (13)$$

$$R_{17} = S_{17} \oplus r_{45} \text{Bdisk} \quad (14)$$

$$V_{14} = (S_{14} - S_{17}) \cap X_0 \quad (15)$$

$$R_{14} = S_{14} - S_{17} \oplus r_{30} \text{Bdisk} \quad (16)$$

$$V_{11} = (S_{11} - S_{14}) \cap X_0 \quad (17)$$

$$R_{11} = S_{11} - S_{14} \oplus r_{27} \text{Bdisk} \quad (18)$$

$V_i (i = 11, 14, 17)$ に対してそれぞれの文字列候補領域についてヒストグラムを求め、各領域のピーク (p_i) と p に隣接する階調値の一つを採用する。(隣接する階調値の採用にあたっては、画素数の多い方を採用) これらの値を R_i のそれぞれ対応する部分にあてはめ、同じ階調値をとる画素を文字列領域 (L_i : 階調値 0 と階調値 i をもつ2値画像) とする。(Fig.10)

2.2.4 ノイズ除去

2.1.6 と同様に L_i で検出された文字列に対して, Conditional Dilation を施し, ノイズを除去し, 文字を整形する。(Fig.11)

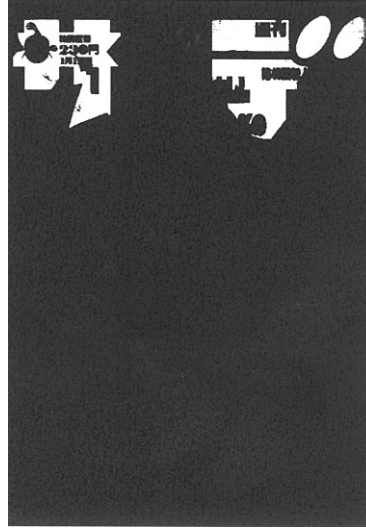


Fig.11

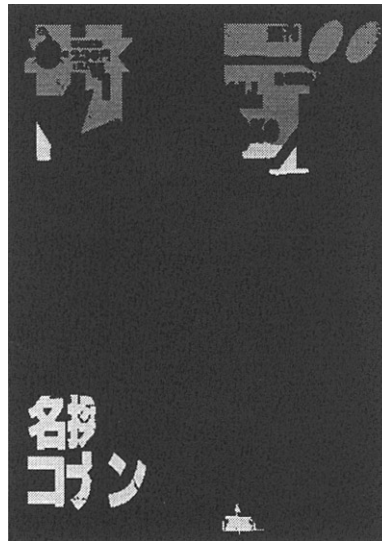


Fig.12 Z_{r2}

$$K'_{17} = L_{17} \circ r_{20} \text{ Bdisk} \quad (19)$$

$$K'_{14} = L_{14} \circ r_{15} \text{ Bdisk} \quad (20)$$

$$K'_{11} = L_{11} \circ r_{11} \text{ Bdisk} \quad (21)$$

$$K_i = K'_i \oplus |_{L_i} \text{ Bsquare} \quad (i = 11, 14, 17) \quad (22)$$

以上の処理で導出された領域画像 K_i がそれぞれ Canny エッジ検出器による文字列抽出結果である。

$$Z'_{r2} = K_{11} \cup K_{14} \cup K_{17} \quad (23)$$

2.3 結果画像の生成

最後に、2.1, 2.2 で得られた各次数における領域の競合を解消するため、 $K_i (1 \leq i \leq 10, i = 11, 14, 17)$ に対して 2.1.5 の Fig.5 の方法を用いて他の次数の文字列領域と重なっている部分があれば削除する。ここから得られた各次数の文字列抽出の論理和をとった画像を最終結果とする。(Fig.13)

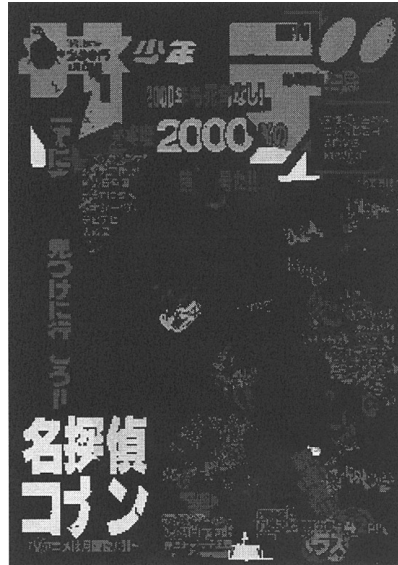


Fig.13 Z_r

$$Z_r = Z'_{r1} \cup Z'_{r2} \quad (24)$$

一つの階調値が抽出された一つの文字列を表す。但し同じ階調値であっても、 $V_i (1 \leq i \leq 10)$ 、 $R_i (i = 11, 14, 17)$ で得られた一つの領域内になければ別の文字列として扱っている。

3. まとめ

Fig.13 に示すような文字列抽出結果が得られた。まず、文字サイズがある程度大きい文字列については良好な結果であることが確認できた。文字サイズの小さな文字列領域については付近の背景部分がノイズとして抽出されている領域が目立つ。これは今回用いた原画像の解像度が 100dpi という低解像度に起因すると考えられる。処理速度の増大を招くため高解像度の画像を用いることは好ましくない。そこで、現在抽出されている文字列領域に関して、その領域のみを高解像度で再サンプリングすることを考えている。また、サンプルとして示した雑誌の他に数種の雑誌の表紙画像についても文字列抽出実験を実施した結果いずれも良好な抽出結果が得られることを確認している。

参考文献

- [1] 顧 力翔, 田中直樹, 金子 豊久, R.M.Haralick, 「表紙画像からの文字領域抽出方式」, 電子情報通信学会論文誌 Vol.J80-D-II No.10 P.2696-2703, 1997/10.
- [2] Petros Maragos, Ronald W. Schafer, Muhammad Akamal Butt, 「MATHEMATICAL MORPHOLOGY AND ITS APPLICATIONS TO IMAGE AND SIGNAL PROCESSING KLUWER ACADEMIC PUBLISHERS (Boston/Dordrecht/London), P.219-226, HUGUËSTALBOT, A MORPHOLOGICAL ALGORITHM FOR LINEAR SEGMENT DETECTION」.
- [3] John Canny, 「A computational Approach to Edge Detection」, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. PAMI-8, NO.6 P679-P698, NOVEMBER 1986.

発 表 資 料

題 名	掲 載 誌 ・ 学 会 名 等	発 表 年 月
Shannon Divergence とモルフォロジーを用いた白線の抽出に関する研究	電子情報通信学会2002総合大会	2002年3月