6Q−6

文書自動登録を目的とした文書セパレート方式

伊勢　広敏*、武田　晴夫*、町田　哲夫*、増崎　秀文**
*（株）日立製作所システム開発研究所
**（株）日立製作所小田原工場

1．まえがき

画像データ、コードデータなど大量の情報を効率良くしかも低コストで、記憶、保存する手段として、光ディスクを用いた電子ファイル装置が広く利用されつつある。電子ファイル装置に文書画像を登録する場合、文書単位にスキャナから画像データを入力し、文書に対応したインデクスを付加する方法が一般的である。したがって、複数の文書を一括入力する時には、文書間を識別する手段が必要となる。

本稿では、セパレータシートを利用して文書間を識別する方法を検討した。ここでは、雑誌、新聞などの文書と容易に識別できるセパレータシートを考案することを目的として、文書に出現する確率が極めて低いマークについて調査した。このマークとしてバーコードパターンを採用し、評価したので報告する。

2．文書セパレート方式

2．1　セパレータシート識別手順

ここで考案した文書セパレート方式は、文書間にセパレータシートを挿入し、セパレータシートを識別することにより、文書を分離するものである。処理概要を図1に示す。

セパレータシートを利用して、文書を分離する手順を以下に示す。

（1）各文書の先頭ページの前にセパレータシートセパレータシートを挿入し、スキャナで一括入力する。

（2）入力画像からセパレータシートを認識し、文書間を識別する。

（3）識別した文書ごとに画像ファイルに格納する。この時、セパレータシートは格納しない。

2．2　セパレータシートの仕様

セパレータシートは、文書と確実に分離できる必要がある。ここでは、文書に出現する確率が低

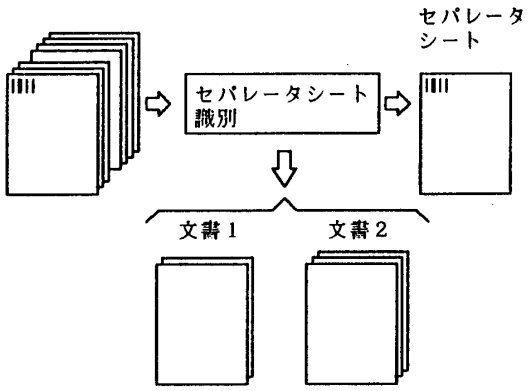

図1　文書セパレート方式

いマークをセパレータシートに付加することを考え、このマークとして、バーコードを利用した。バーコードとは、バーとスペースのそれぞれの幅の長短により、キャラクタを表現するものである。

表1　キャラクタの出現確率

キャラクタ	右側		左側	計
	O	E	E	
0	1108	282	713	2108
1	806	102	354	1262
2	139	86	322	547
3	591	2550	1616	4752
4	358	626	453	1437
5	1553	521	477	2551
6	715	4202	1340	6257
7	327	2451	1036	3814
8	306	2016	641	2963
9	309	275	776	1360
計	6212	13111	7728	27051

チェック対象数：3894912×8

Document Separating for Automatic Registration
1:Hirotoshi ISE,1:Haruo TAKEDA,1:Tetsuo MACHIDA,2:Hidehumi MASUZAKI
1:Systems Development Laboratory,Hitachi, Ltd.,2:Odawara Works,Hitachi, Ltd.

ここで使用したバーコードはＪＩＳ規格であるＪＡＮコードであり、バー２本とスペース２本で１キャラクタを構成し、０～９までの数字のみを表す。ＪＡＮコードの各キャラクタがＣＣＩＴＴ標準原稿（Ｎｏ．１～８）に出現する頻度分布を表１に示す。

表１より、以下のことがわかる。

（１）キャラクタの出現確率は、８．７×１０$^{-4}$である。

（２）出現確率は、キャラクタによりバラツキがあり、「２」、「９」の出現確率が小さい。

キャラクタがｎ桁連続して出現する確率について調査結果を表２に示す。

表２　連続桁数と出現確率

連続する桁数	出現確率
1	7.1×10^{-4}
2	7.2×10^{-6}
3	7.6×10^{-8}
4	1.5×10^{-9}
5	3.3×10^{-7}
6	6.5×10^{-8}
7	3.2×10^{-9}
8	0.0

表２より、キャラクタがｎ桁連続して出現する確率ｐ（ｎ）は、以下の式に従う。

ｐ（ｎ）＝７．１×１０$^{-4}$×０．１５

なお、ＪＡＮ標準バージョンは１６桁のキャラクタ、ＪＡＮ短縮バージョンは１１桁のキャラクタで構成されている。したがって、これらのコードが文書に出現する確率は、

標準バージョン：ｐ（１６）＝４．６×１０$^{-17}$

短縮バージョン：ｐ（１１）＝４．０×１０$^{-12}$

となる。

キャラクタを特定した場合には、短縮バージョンでも、出現確率は、１０$^{-20}$以下である。また、表１より、単独のキャラクタとしては、「２」の出現確率が最小である。したがって、全てのキャラクタを「２」で構成したＪＡＮ短縮バージョンをセパレータシート用のバーコードとした。このバーコードが文書に出現する確率は、１０$^{-25}$以下であり、セパレータシートが文書に誤認識される確率は極めて低いといえる。

３．セパレータシート認識方式

スキャナで一括入力された文書から、前述した仕様のバーコードが付加されたセパレータシートを認識する方式を示す。ここに示すセパレータシート認識方式は、メモリ上にある入力画像を２次元的に処理することによりバーコード認識するものである。

ここで評価したバーコード認識方式は、バーコードが垂直方向（ライン間）に相関性を持つ点に注目している。この点を利用する認識方式としては、垂直方向の周辺分布をとる方式、ライン単位に認識し頻度分布を調べる方法などが考えられる。ここでは、識別時間および画像の傾きを考慮し、隣接するラインを演算処理してバーコード認識した。具体的には、ライン間の論理和によりかすれを補正し、論理積によりノイズ除去した画像に対してパターンマッチングすることによりバーコードを認識した。このバーコード認識方式を評価するため、バーコード画像に白色雑音を重畳し、重畳した雑音の割合とバーコード認識率の関係を調査した。調査結果を図２に示す。

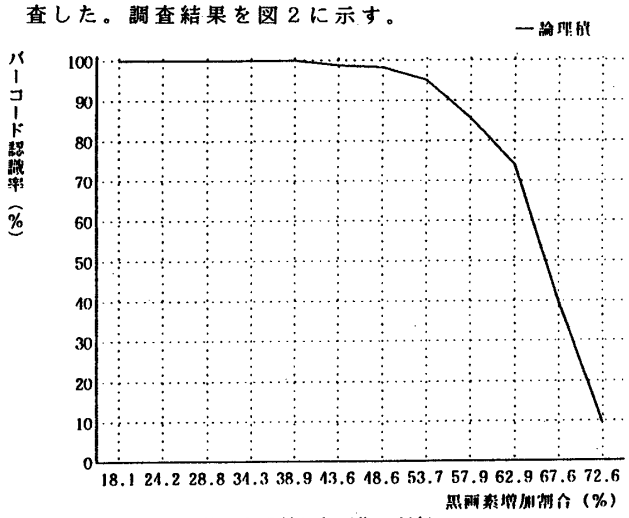

図２　白色雑音を重畳した画像の評価

上記の方式でＣＣＩＴＴ標準原稿を識別した場合、「文書」、「セパレータシート」、「識別不可」に認識される割合は、

ｉ）文書：９９．９９９９９９％以上

ii）セパレータシート：２．６×１０$^{-23}$％以下

iii）識別不可：８．０×１０$^{-7}$％

となる。

４．あとがき

セパレータシートを利用して文書間を識別する方式について報告した。バーコードパターンは、文書に出現する確率が小さいことを確認した。特に、「２」を連続したパターンは出現確率が極めて小さく、１０$^{-25}$以下である。また、セパレータシートを本認識方式により識別した場合、文書がセパレータシートに誤認識される割合は、２．６×１０$^{-23}$％以下であり、セパレータシートとして実用可能である。