1P-3 文書認識処理の高速化を指向した専用ハードウェアの検討

岩城　修　　木田 博巳

NTT　電気通信研究所

1. まえがき

文字、図形、イメージが混在したマルチメディア文書から各メディアを分離抽出し、後段の処理が容易なようにメディア変換する機能の実現が望まれている(1)。文書画像の処理は膨大なデータを扱うことから、処理の高速化が必須である。これまでアルゴリズムの立場から並列処理等による処理の高速化の検討を行ってきた(2)。本稿では文書認識処理において、パイプライン処理、並列処理をハードウェアで実現した場合の性能評価実験を行った結果について報告する。

2. 文書認識処理

文書認識処理は、前処理、領域抽出処理、テキスト領域認識処理、表領域認識処理、図領域認識処理に大別される(3)。文書認識処理の流れを図1に示す。前処理部では文書画像の復号化、雑音除去を行い、領域抽出に必要となる周辺分布、黒連結、線分の各特徴を抽出する。これらの特徴に基づき、領域抽出部ではテキスト/表/図領域を抽出する。テキスト領域認識部では一文字の抽出と文字サイズの正規化(拡大/縮小)を行って文字認識する。表領域認識部では文字の認識と表罫線の認識を行う。図領域認識部では近接線密度特徴を抽出して文字/図形分離を行い、文字の認識と図形領域の画像符号化を行う。

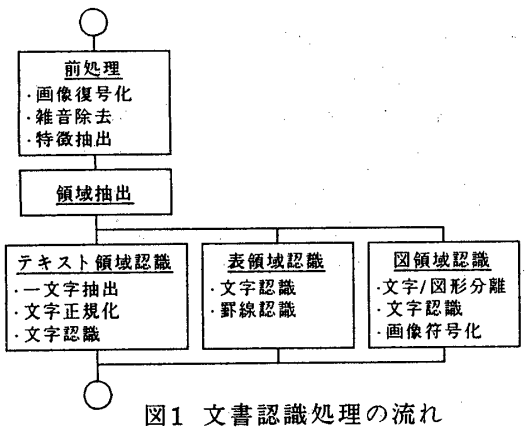

図1 文書認識処理の流れ

3. 処理の高速化

① パイプライン処理

文書認識処理では特徴抽出等パターンレベルの処理が多く、これらの処理をハードウェア化して処理時間の短縮を図る。また入力文書画像に対し複数のパターンレベルの処理を連続的に行う場合があり、これらはパイプライン処理によりメモリアクセス等を軽減し高速化を図る。

② 並列処理

一文書の認識時間(ターンアラウンドタイム)及び単位時間内に認識できる文書数(スループット)を向上させるため並列処理を行う。並列処理には文書内で行う場合と複数の文書に対して行う場合がある。前者の例には文字の認識と並行して次行の文字の抽出や拡大/縮小を行う場合があり、後者はイメージメモリ上に複数の文書画像を配置しマルチに処理する場合が処理の高速化に有効である。

4. ハードウェアの構成

ハードウェアの構成を図2に示す。文書認識処理の流れを制御する基本部、文書画像を高速に処理するメディア処理部及び漢字OCRから構成され、メディア処理部はさらにメディア処理プロセッサ、イメージメモリのほか、画像符号/復号化、雑音除去、周辺分布特徴抽出、黒連結特徴抽出、線分抽出、近接線密度特徴抽出、拡大/縮小の各画像処理専用ハードウェアから構成される。各画像処理

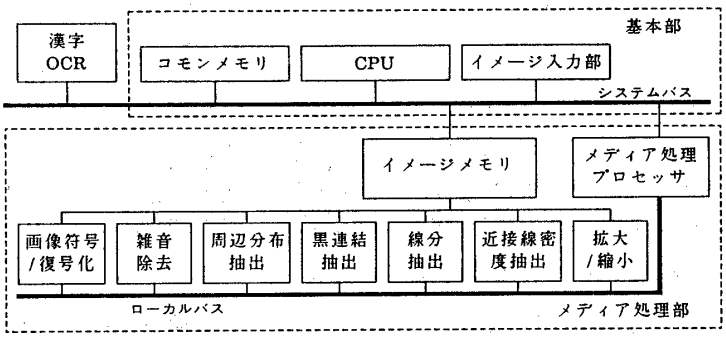

図2 ハードウェアの構成

A high-speed image processing hardware for document recognition

Osamu IWAKI　Hiromi KIDA

NTT Electrical Communications Laboratories

ハードウェアはイメージメモリ上に配置された文書画像に対し処理を行い、複数の処理を連続して行うパイプライン処理や、同時に行う並列処理が可能である。またCPUとメディア処理プロセッサによる並列処理も可能である。

5. 性能評価実験

5.1 パイプライン処理の評価実験

前処理部で行う画像の復号化、雑音除去、周辺分布・黒連結・線分の各特徴抽出をパイプライン処理する場合と個別に行う場合について処理時間を測定した。その結果、個別に行った場合はイメージメモリのリード/ライトが占める割合が多く、雑音除去が最も処理時間を要した。パイプライン処理を行った場合は雑音除去の処理時間で終了し、処理の高速化が図れることが分かった。

5.2 並列処理の評価実験

試作した文書認識装置を用いて、前処理部、領域抽出部、テキスト領域認識部、表領域認識部、図領域認識部の各モジュール毎の処理時間を測定した結果を表1に示す。表1では前処理時間を1として画素当たりの処理時間を相対値で示している。その結果、テキスト領域の処理時間が最も長く、その内部の処理時間は図3に示す負荷バランスとなっている。図3の各処理は独立のハードウェアで行われることから、テキスト領域の認識時間はOCRの処理速度に支配されている。

表1 文書認識処理の処理時間

処理内容	処理時間(相対値)
前処理	1.0
領域抽出	0.7
テキスト領域認識	5.5
表領域認識	3.5
図領域認識	3.0

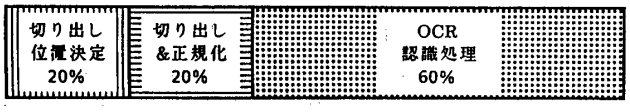

図3 テキスト領域認識時の負荷バランス

文書全体の認識時間は文書中に含まれる各領域の混在比等に左右される。ここでは一つの典型として表2に示す文書構成を想定し、試作装置で並列処理を行った場合の処理能力を求めた。図4は装置内で多重処理する文書数(マルチ度)をパラメータに、ターンアラウンドタイム及びスループットを

表2 文書の構成

処理内容	処理時間(相対値)
テキスト領域	50% (1250文字)
図表領域	10% (150文字、30図形要素)
イメージ領域	15%
余白	25%

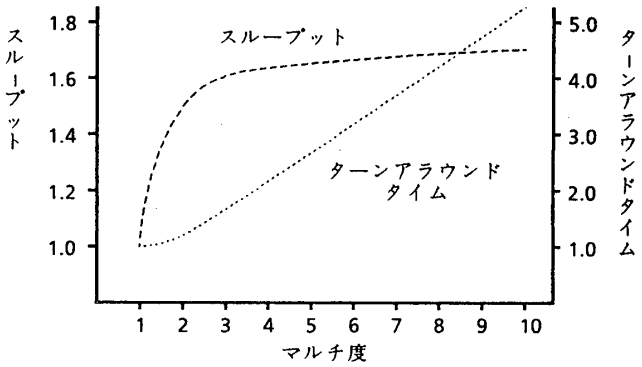

図4 システムの性能

マルチ度1の場合との相対比で示している。マルチ度を上げるに従い、各モジュールの稼動率が高まりスループットが向上する。マルチ度3でターンアラウンドタイムは約2倍となるものの、スループットは約1.6倍に向上できる。しかし特定のモジュールの稼動率が100%に近づけば、それ以上マルチ度を上げてもスループットの向上は望めない。本試作装置の場合、OCRがボトルネックとなっており、マルチ度4以上でスループットの向上は飽和傾向を示す。

6. むすび

本報告では、文書認識機能を実現するための専用ハードウェアを用い、パイプライン処理、並列処理に関し性能評価実験を行った結果を示した。今後、本システムの認識性能について評価を行う。

最後に本検討を進めるに当たり御指導頂いた川田分散処理装置研究室長、荒川主幹研究員、並びに熱心な御討論を頂いた分散処理装置研究室の諸氏に感謝する。

参考文献

(1) 木田他、情処学会「LAN/マルチメディアの応用と分散処理」シンポジウム論文集、pp133~140(1984)
(2) 岩城他、昭60信学全大、1544(1985)
(3) 岩城他、昭60信学会情報・システム部門全大、S4-5(1985)