Mem. Fac. Sci. Eng. Shimane Univ.
Series A
33, pp. 51-62 (1999)

利用者からみたOCR誤認特性

佐藤　匡正*・岸本　頼紀**
*島根大学総合理工学部　数理・情報システム学科
**島根大学大学院理学研究科　計算機科学専攻

Error Properties in Character Recognition of OCR in Utilization

Tadamasa SATOU
Department of Mathematics and Computer Science, Interdisciplinary Faculty of Science and Engineering Shimane University
Yorinori KISHIMOTO
Department of Computer Science, Graduate School of Science Shimane University

Abstract

The character recognition through OCR (optical character reader) is possible to include several errors. The errors are remarkable in recognizing such sentence as include various font types, along with errors being due to stain or wrinkle. To wittily OCR it is important to hold the factors of recognition errors. This paper presents OCR recognition error (trends by analyzing experimental data in recognizing some kinds of manuscripts. The manuscripts to be read contain low quality dot-matrix prints and book pages with regular types, with program codes and mathematical formula, and with both regular and ruby (tiny) types. In the analysis errors are classified into three types to clarify that the recognition error trends both in each type and being due to ruby types.

要　　約

OCRによる文字認識は，誤認の可能性がある．誤認は，資料の皺や汚れなどによるものもあるが，異なる寸法の字体が混在する文では特に顕著である．OCRを利用する場合，こうした誤認がどのような状況で生ずるかを把握することが大切である．そこで，誤認の性質を把握することを目的として様々な資料における誤認の状況を調べる．資料としては，印字品質の低いドットプリンタで出力された文，数式やコードを含む文，縦書きの文，ルビをもつ文の4種類を使用した．この調査では，誤認を3種類に分類し，それぞれの分類における誤認の傾向，ルビ文字に起因すると推測される誤認の傾向について分析した．

1. はじめに

古くからの文字入力手段として，スキャナ装置とPCソフトウェアの組み合わせによる光

学文字読取りOCRがある．製造会社によれば，こうしたシステムの読取りには様々な方法がとられており，この性能は正認率で1に近いとされている．しかし，この認識には誤認の問題がある．誤認は資料の汚れや皺，字体という物理的な要因によるものもあるが，文字寸法の異なる文字が混在すると，文の読取り誤りが顕著に表れるという経験をする．自然言語の研究などで辞書の作成や分析にコーパスとして大量の資料を必要とする場合などではOCRが効果的に活用できる．ところがこの場合，正認率は1でなければ全ての認識結果についての確認が必要となる．これには相応の作業を伴うことになる．そして，認識が少なくなればなるほど，また誤りが尤もらしければらしいほど見逃す可能性は高く作業は難しくなる．OCRを利用する立場からは，こうした確認作業において誤認が検出しやすいと助かる．そのためには，誤認の生ずる状況が把握できればよいし，また新たな認識の方法を考えるヒントにもなるであろう．本稿では，横書き文，数式やコーディングを含む横書き文，縦書き活字文，縦書きルビ付活字文について，使用者の立場からみたOCRでの文字認識特性を比較・分析する．

2. 分 析 方 法

誤認についての分析をするにあたっての考え方及び分析方法や計測の条件などについて述べる．

(1) 分析の考え方

OCRの識字性能を決める要因は複雑である．スキャナの読取り精度，原稿の状況，ソフトウェアで用いているアルゴリズムや性能などが念頭に浮かぶ．従って，OCRの性能や特性を一般化して論ずるためには，こうした要因を把握せねばならない．

しかし，利用者の立場から考えると事情は異なる．費用の面から，理想的な資材やソフトウェアの購入は制限されるし，入手したソフトウェアの機能の改良や追加についても全く閉ざされている（著作権法などによる制限）．従って，利用者の行うことは，現行システムの特性を把握して，その欠点を補うような効果的な利用法を確立することである．こうした考えからOCRシステムの識字における誤認の特性について分析を行う．

従って，本分析の目的はいわゆる複数社の製品を俎上に乗せたベンチマーク試験や製品としての評価を目的とするものでない．こうした主旨からスキャナ及びOCRソフトウェアの製造社の社名は明記しない．

(2) 誤認の分類

字の認識誤りについての考え方を整理する．

■誤認

OCRの認識した字を読取る原稿（正本と言う）の字と対応づけて比較するとき，一致する場合を「正認」，不一致する場合を「誤認」と言うことにする．この誤認は現象面と原因面に分けて捉えることができる．

■現象面の誤認

誤認を現象面から見る．正本を基準として誤認を見ると，この現象は，字の欠けている

もの，余分な字の加わっているもの，字が別のものに化けているものの三種類が考えられる．これらをそれぞれ「欠」「余」「化」と言うことにする．「欠」は正本の字が識字されず，欠けている状態である．「余」は正本に無い字が識字され，字の加わった状況を言う．「化」は互いに対応している文字の異なるものである．

■原因面の誤認

上の三種類の誤認の生ずる原因を考えてみる．「欠」は字の写りが薄く汚れとみなされるのであろうし，「余」字現象では，原稿のゴミや読取り台のキズを逆に字と誤る．また，「化」では，類似した字形を別の字と誤る（識字誤り）ことや，字の句切り方（句切り誤り）を誤ることである．

■句切誤り

よく出会う誤認は「化」なので，これに着目する．「化」誤認は大別すると，上で述べた「識字誤り」と「句切り誤り」がある．字数の対応で言えば，正本の字を1としたとき，前者は1：1対応であるし，後者は1：nである．但し，n−1字が正本の字の要する（部首など）に対応しているものを言う．例えば，「誤」を「言」と「呉」の2文字に識字するように辺と旁を別個の文字とするものである．

■ルビ誤り

ルビを含む文については，句切り誤りの更に複雑化した状況が生ずる．つまり，ルビ文字と地文の文字が一体化する可能性が考えられる．

(3) 分析システム

比較の確度の向上及び，統計処理の容易化をねらいとして，正本と認識結果を比較するツール（知的比較）を開発して分析を行う．分析システムの構成を図1に示す．原稿をスキャナを経由して，OCRソフトウェアに入力し，認識結果を得る．この複写を人手によって修正し，正本ファイルを作成する．これとOCR認識結果との一致具合を比較する．この一致比較（知的比較）は正本の行の基準にして1字ごとの「化」「欠」「余」現象を識別できる仕様をもつプログラムである．このプログラムの概要を付録1に示す．

(4) 読取り条件

認識の性能は読取る原稿の種別によって変化すると考えられるので，代表的な原稿として，次の4種類を用意する．なお，スキャナの読取り解像度は600 dpiである．

① プリンタ出力（ドット字形）した横書き文

B5版，10ページ，284行の量である．字体は明朝体，約13ポイントで，全て同一の大きさである．読取り原稿は，ドット印刷機による出力をゼロックス機で複写したものである．

② 数式やコーディングを含む横書き文

文献2の12ページ，354行である．スキャナの読取り原稿は文献2をゼロックス機で複写したもの．

③ 縦書き活字文

文献3の20ページ，355行である．読取り原稿は書籍を直接使用する．ルビ文字取り除いたものを原稿とする．

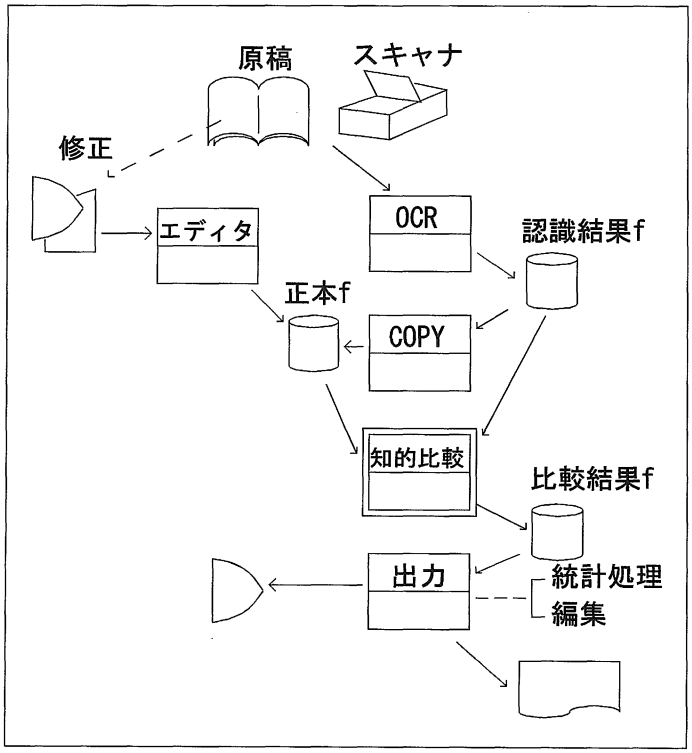


図1 分析システムの構成

④ ルビ付き縦書き活字文

文献3の20ページ，524行である（ただし，ルビ文字の行は1行とする）．つまり上の③の読取り原稿をそのまま使用する．

3. 分　　析

3.1 資料全体の認識状況

資料全体の認識状況を表1に示す．ただし，資料②④においては，ルビ文字および半角文字も1文字と数える．正認率を正認字数/正本字数として示す．

3.2 行単位の正認率

正認率は行によって異なる．各資料の行による正認率の分布を表2に示す．ただし，資料④においてはルビ文字部の欠字が多く見られ，これを分析しても無意味と考え，ルビ文字の行は1行と数えず，ルビ文字部とそれのふられている本文文字を合わせて1行として数

表1　認識状況

	正本字数	正認字数	誤認字数	正認率
資料①	7765	7570	1022	0.97
資料②	9862	9788	833	0.99
資料③	9944	9872	156	0.99
資料④	11241	10418	1595	0.92

表2　行正認率の分布

	出現比率			
	資料①	資料②	資料③	資料④
0～0.6	9.5%	2.8%	0.6%	10.3%
0.6～0.8	16.2	3.7	0.9	24.1
0.8～1.0	54.6	49.1	27.4	39.7
=1.0	19.7	44.4	71.0	25.9

表3　誤認の状況

分類	比率（%）			
	資料①	資料②	資料③	資料④
欠	3.7	1.4	4.7	13.4
余	12.0	0	2.0	4.3
化	84.3	98.6	93.3	82.3

える．

3.3　誤認の状況

誤認の集計結果を表3にまとめる．これにより誤認の大部分は化字であるといえる．また資料④では化字はルビ文字で生じていることが多い．これは，ルビ文字を本文文字の一部として誤認し，本文文字とルビ文字を合わせて1文字として認識していることによる．

3.4　化字認識

(1)　識字誤りの状況

①　文字の種類による分類

文字の大きさによって識字誤りは変化する．資料②において，半角文字の正認率は

75.4%であり，半角文字を除いた，全角文字のみの正認率は97.5%である．また，資料④においてルビをふられている文字のみに着目した場合，正認率は31.3%であり，同じ資料②でルビを取り除いた資料（スキャナでデータを取り込んだ後，データ上で修正した）の正認率は97.5%である．これにより，幾つかの大きさの文字が混在する文を認識する場合は，識字率が著しく下がることがうかがえる．

② 傾向

■記号

全体を通して「、」や「；」等の記号文字が識字誤りの32.4%を占める．これは，誤認されやすいことを意味している．

■特定の文字

同様に識字誤りを生じやすい特定の文字も存在する．全体を通して「プ」「ガ」のように，カタカナの濁音・半濁音文字は「フ」「カ」のように濁音・半濁音を取り除いた識字誤りを生じやすい．

(2) 句切誤りの状況

① 文字の種類による分類

句切り誤りは半角文字で19.4%全角文字では80.5%である．また資料④においては，ルビ文字を本文文字とまとめて１文字として誤認している傾向があり，ルビを含む文字における句切誤りは，資料④における誤認の50%を占めている．これは，半角文字やルビ文字のように大きさの異なる文字の扱いが難しいことを意味している．

② 句切誤りの分類

文字の認識での文字の句切は，基本的には次の３つに分類できる．

1：一字を複数文字に句切る

2：複数文字を一字に句切る

3：文字列を適当に句切る

表４に，この句切り誤りによる出現数を分類して示す．資料①②の場合は（「語」→「言吾」）等の偏と旁に別れる句切誤りが多いのに対して，資料③では句切り誤りは殆ど見られない．資料④ではルビ文字を本文文字と合体して１文字として認識する句切り誤りが多く見られる．

表４ 句切りの分類と誤り出現状況

句切の分類	出現比率（%）					
	資料①	資料②		資料③	資料④	
		全角	半角		本文	ルビ
一字を複数文字に	52.3	27.8	4.2	91.6	28.0	62.5
複数文字を一字に	14.0	16.7	87.4	4.2	52.0	12.5
文字列を混合して	33.6	56.5	8.3	4.2	20.0	25.0

4. ルビ文字の影響

ルビ文字は，読者の便に配慮して難解な漢字の読みを補助するためのもので，通常漢字の横に置かれる小型の字体（ルビ活字（ruby））が用いられる．文字の認識においては漢字とルビを一体化した文字として認識する可能性がある．この誤認は特別なので，項を設けて述べる．

(1) 認識におけるルビ文字の影響

ルビ文字は識字に影響を与える．図2に資料③と資料④の1ページにおける正認率と出現頻度のグラフを示す．グラフの横軸は正認率であり，縦軸がその正認率の出現頻度である．

(2) 文字構造に見られる句切誤り

①漢字

ルビ文字を含む文の誤認には，構造上の特性が見られる．表5にルビを含む文字の構

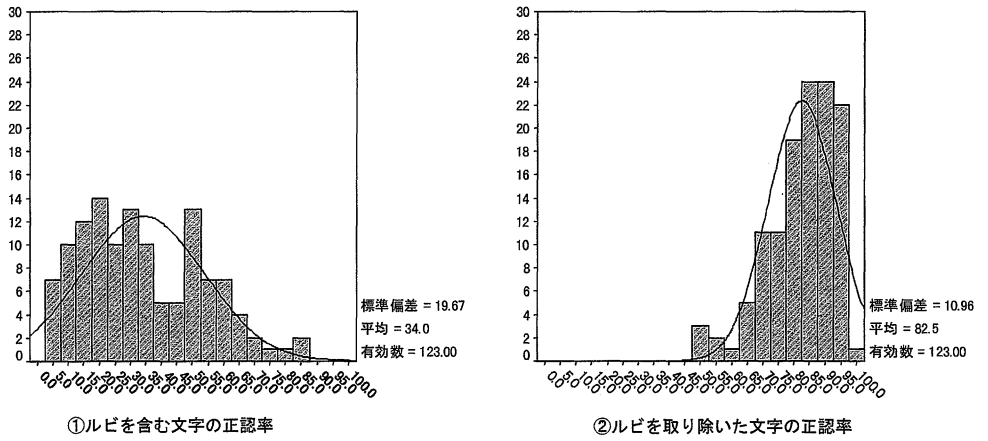


図2　ルビ文字の正認率への影響

表5　文字構造の特性

被認文字	認識結果（%）			
	ひらがな	1 塊	2 塊	3 塊
ひらがな	82.6	3.3	12.4	1.7
1 塊	6.5	7.2	83.4	2.9
2 塊	0	2.5	60.5	37.0
3 塊	0	7.1	86.9	6.0

造の特性を示す．被認文字が１つの塊，例えば「上」「今」のような場合，認識結果文字が２塊に誤認される割合は83.4％である．また，被認文字構造が２塊である場合は，認識結果文字のうち60.5％が２塊の構造をもつ誤認であり，37.0％が「粥」「蹴」のような，偏と旁の間にもう１つ部をもつ３塊に誤認されている．

②ルビ

地文とルビ文字についての誤認の組み合わせを示したものを表６に示す．全体の誤認のうち，46.0％が「ルビは認識せず・本文は誤認識」している．これはルビ文字が本文文字の部首を成す偏と旁と誤認されるためと推定される．

③ひらがな

ひらがなに見られる句切り誤りを表７に示す．被認識文字がひらがなの場合，被認文字に関わらず「が」と認識している割合が高い（47.1％）．これは，ひらがながルビをも

表６ ルビ付文字の認識状況

認識		比率（％）
地文	ルビ	
○	○	31.1
×	○	1.2
○	‖	8.3
×	‖	46.0
○	×	4.9
×	×	0.7
句切誤り	まとめ	2.9
	分解	3.1
その他		1.7

表７ ひらがなの誤認結果

出力結果	割合（％）
い	18.2
か	7.4
その他（ひらがな）	5.8
が	47.1
その他（濁音・半濁音）	4.1
その他（漢字）	14.0

つ場合は，その直上の漢字に対するルビが下にはみ出している場合が大半であり，ルビを「が」の右にある「゛」の部分と誤認しているためと推定される．

5. 結　　論

利用者からみたOCR誤認特性を分析した．その結果，次のことが分かった．

1) 縦書き，横書において正認率に差は見られない．
2) ルビ文字や半角文字を含む文は誤認が起こりやすい．
3) 誤認を「欠余化」字に分けたとき，欠字が生じやすいのは記号や空白である．
4) 余字は，行頭，行末，または行と行の間で生じ，記号に認識されることが多い．
5) 化字は，識字誤りと句切り誤りに分けた場合，識字誤りが約2/3を占める．
6) 識字誤りは，記号や半角文字に多い．また特定の文字に片寄る傾向が見受けられる．
7) 句切誤りでは，横書きではその半数以上が一字を複数文字に句切る．

縦書きではほとんど見られないが，ルビ文字を含む場合は，ルビ文字を本文文字と合体して誤認する句切誤りを多く生じる．特定の文字について繰り返し生ずる傾向が伺える．

参 考 文 献

1) 佐藤匡正「利用者からみたOCR認識誤りの分析」電気・情報関連学会中国支部第49回連合大会，講演論文集（1998）
2) John R. Leviue/Tony Mason/Doug Brown; lex & yacc; O'Reilly & Associates Inc. 村上　列訳　アスキー出版局（1994）
3) 荒俣　宏「帝都物語8」角川書店

付録．知的比較システム

1. あらまし

本システムは，正本ファイルと被験（認識結果）ファイルの対応するレコードの文字列を比較して，化字，余字，欠字の指摘を支援する機能をもつ．

2. 機能仕様の概要

(1) 知的比較

2つのファイルに対して，後述する4. 比較仕様に基づいて比較を行い，その結果をファイルに出力する．この比較を「知的比較」という．

(2) ファイル群の自動比較

指定されたディレクトリ名配下の全ての基準及び被験ファイルについて知的比較を自動的に実施し，その全てをファイルに保存する．

(3) ディレクトリ名及び作業用ドライヴ番号の指定

比較対象のファイルの置いているディレクトリ名及び作業用ドライブ番号が指定できる．

(4) 出力結果

比較結果は全て蓄積に指定されたディレクトリ名の下に作成したファイルに出力する．

3. 操作仕様の概要

(1) 起動

次のように打鍵すると動作が始まる．

readCk〈比較ファイル郡のあるディレクトルリ〉〈作業用ドライヴ名〉

ここで，〈比較ファイル郡のあるディレクトリ名で，最後に“¥”を欠かさない．指定しない場合は“a：¥”とする．作業用ドライヴ名は英1時と“：”で指定する．指定を省略すると“g：”と解釈する．

(2) 指定の確認

起動時の指定を確認のために表示する．中止するときは「ctrl+c」による．

(3) 中途の表示

処理中中途には画面に中間結果が表示され，かつ介入操作はできない．

(4) 結果の参照

〈指定したディレクトリ名〉r.acc 名を持つファイルをエディタで参照する．

4. 比較仕様の概要

(1) 比較対象

同一本体名をもつ つのファイルを比較する．ファイル名本体をfとしたとき，f.txR と f.txC の2つのファイルから同一レコード番号のレコード内全ての文字列を比較する．

(2) 基準文字列

f.txR からの文字列を基準として，f.txC からのを被験文字列とする．

(3) 比較結果と出力

比較の結果は，一致，化字，余字，欠字の4種類がある．

①一致の場合は被験文字列を“_”に変えて出力する．

②化字は同一長の異なる部分列をいう．披験文字列を出力する．

③余字は基準に含まれていない部分列をいう．基準には空白をその余字列長だけ補って出力し，被験字の方はそのまま出力する．

④欠字は基準にはあるが，被験にない文字列をいう．基準はそのまま出力し，被験はその欠字列長だけの空白を補って出力する．

以上．

(4) 留意事項

①レコードの食い違い

レコードを単位として比較をするので，基準と被験についてレコードの食い違いのある場合は対応していない．

②比較文字列に同一部分文字列のある場合

現状では配慮していないため結果が乱れる．

5. 構成仕様

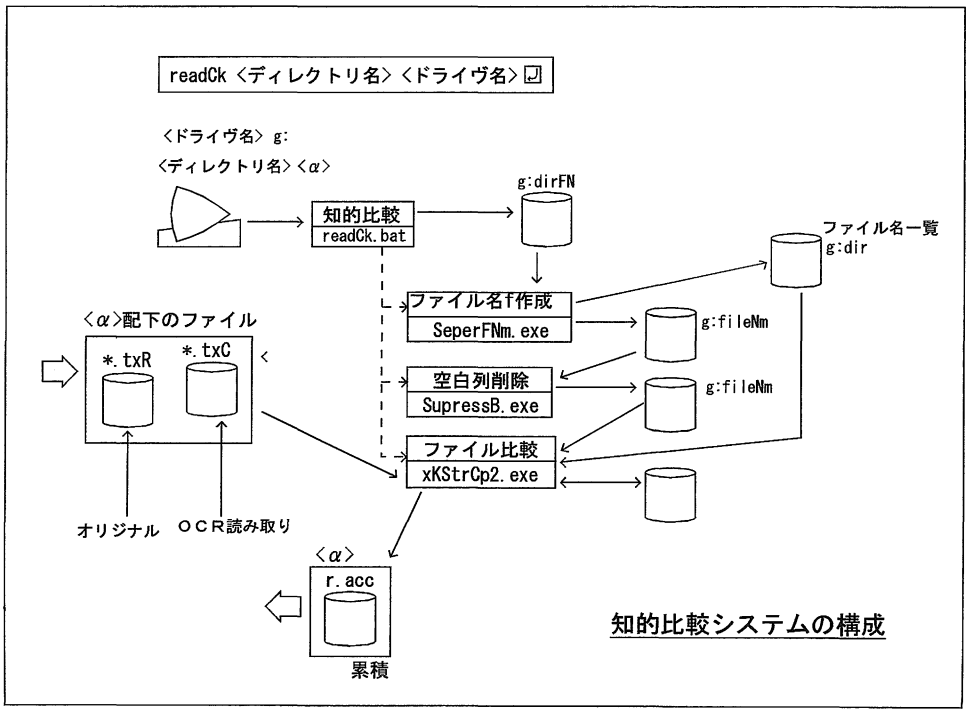
付図1に示すように2つのサブシステムである「ファイル名f作成」及び「ファイル比較」から成る．2つのシステムを制御しているのは readCk.bat である．なお，「空白列削除」はオプションである．

5.1 ファイル名f作成サブシステム

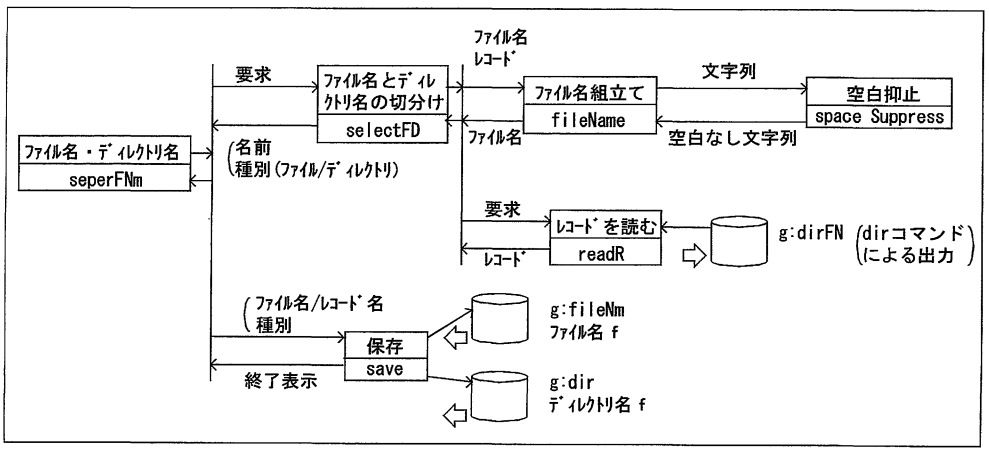
付図2に示す構成をもつ．知的比較で用いるファイル名の組を作成する．

5.2 ファイル比較サブシステム

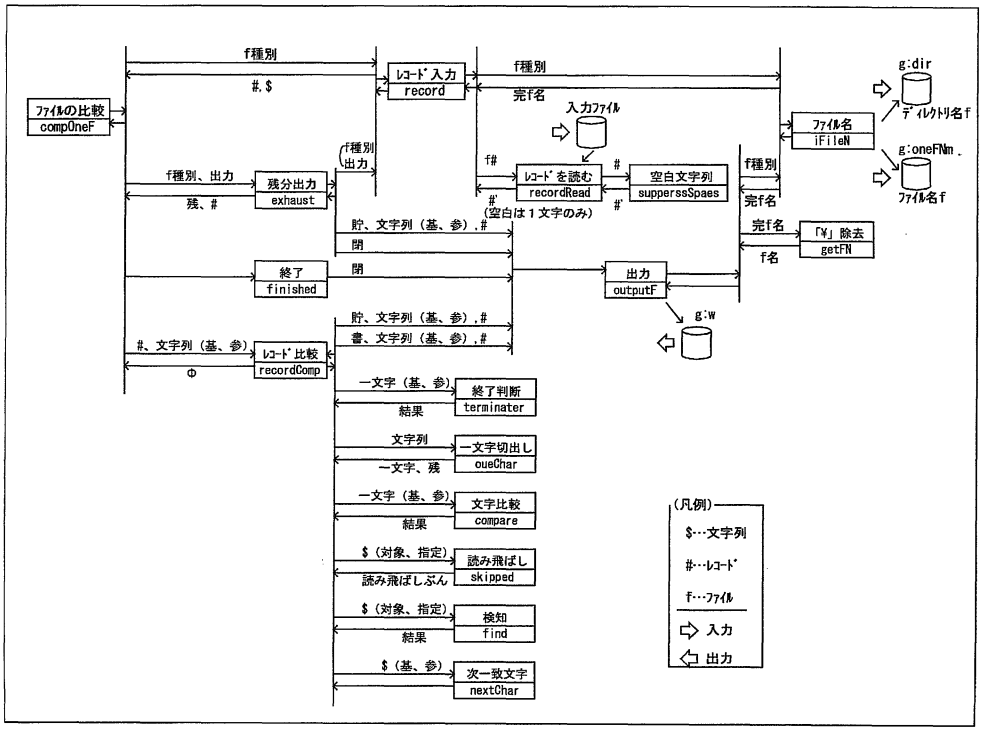
正本ファイルと被験ファイルの組のレコードを対応させて読み，文字列と比較するが，その構成を機能関連図として付図3に示す．



付図1　知的比較システムの構成



付図2　ファイル名及びディレクトリ名



付図3　ファイル比較サブシステム