文字認識を利用したホームページ自動作成

5Q－7

岡田　康裕　　依田　文夫
三菱電機株式会社　情報技術総合研究所

1．はじめに

インターネットの普及により、WWW（World Wide Web）を用いた情報の収集と発信が頻繁に行われるようになった。この情報伝達の媒体には、HTML（Hyper Text Mark-up Language）で記述したホームページが用いられる。そのため、情報を発信するにはHTMLやパソコンに関する専門知識が必要であり、上記の専門知識を有しない情報提供者でも、ホームページを簡易に作成できる手法が求められている。

従来、ホームページを簡易に作成する手法として、手書きで記入された帳票をFAXから送信し、これを情報センタ側で自動認識することによりホームページを作成する手法が提案されている[1][2]が、グラフィックも含めた多彩なホームページを作成することはできなかった。このため、帳票に記入された文字とマークを認識した結果に、事前に用意された文字列と画像情報とを組合せることにより、カラー画像や動画を含んだ多彩なホームページを簡単に作成する手法について検討したので報告する。

2．基本構成

2．1　システム概要

ホームページ自動作成を行うシステムの概要を図1に示す。情報提供者は、情報センタであらかじめ作成した帳票に、ホームページに記載したい内容を手書きで記入し、帳票をFAX送信あるいは郵送により情報センタに返送する。情報センタでは帳票内の文字及びマークを自動認識し、認識結果をHTML形式に変換する。

2．2　ホームページ自動作成処理の流れ

図2に処理フローを示す。FAXまたはスキャナから入力された画像と、帳票の構造を記述した帳票フォーマットとを照合した後、帳票内の文字及びマークの認識を行う。次に、HTMLへの出力形式を後述するスクリプト言語で定義した出力フォーマットに文字及びマークの認識結果を当てはめ、帳票に記入された内容とあらかじめ登録された文字列及び画像を組合せて帳票内の情報をHTML形式に変換し、ホームページを作成する。以下、HTML変換について詳細に説明する。

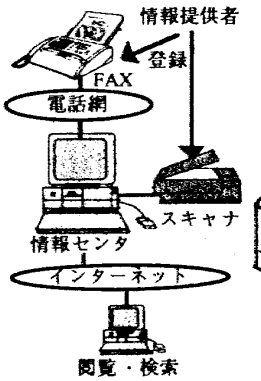


図1　システム概要

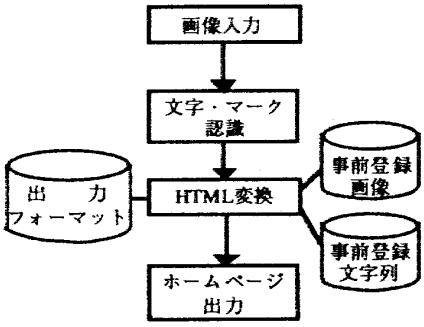


図2　処理フロー

3．HTML自動変換

3．1　HTML自動変換スクリプト言語

文字・マークの認識結果と、あらかじめ登録された文字列及び画像を組合せたホームページを自動作成するため、表1に示すスクリプト言語を導入する。HTML変換はスクリプト言語で記述した帳票固有の出力フォーマットをもとに自動的に実行される。

表1　スクリプト言語仕様

記述形態	意　味
<#N#>	フィールドNの内容を出力する。文字またはマークが記入されるフィールドに対しては、認識結果のコードを出力し、画像が記入されるフィールドに対しては、フィールド内の画像を格納したファイル名を出力する。
#IF<N> ～ #ENDIF	フィールドN内のデータの有無に従って#IFと#ENDIFの間に記述された内容を出力する。Nは、1つのフィールドまたは複数のフィールドを|(or)または&(and)で繋いだもの
<<N>>	フィールドNにマークが付されている場合に、マークに対応する固定文を出力する。

Automatic Homepage Production with Character Recognition
Yasuhiro Okada, Fumio Yoda
Mitsubishi Electric Corporation
Information Technology R&D Center
5-1-1, Ofuna, Kamakura, Kanagawa, 247, Japan

3．2　HTML自動変換機能

3．1節で示したスクリプト言語により、下記の機能が実現できる。図3に、(a)入力画像、(b)スクリプト言語により記述された出力フォーマット、(c)変換後出力されるHTMLファイルの内容、及び(d)画面出力の例を示す。

（1）文字認識結果の出力

帳票読取の結果、得られた文字認識結果を<#N#>記述子により、HTMLの文字列として出力する。

（2）マークの検出と対応文字列の展開

帳票読取の結果、得られたマーク認識結果に従い、文字列を出力する。図3の例では、(b)出力フォーマットに記述された『<<17|18|20>>』により、3つの選択肢の中からマークが付された文字列を出力する。

（3）固定文の出力

ホームページを作成する上で認識する必要がない固定された文字列については、あらかじめ出力フォーマットに直接記述しておき、HTMLの文字列として出力する。図3の例では、(b)出力フォーマットの『商品名』などの記述がこれに該当する。

（4）帳票に記載した画像の出力

帳票読取の結果、帳票内から抽出した画像情報を<#N#>記述子によりHTMLの画像として出力する。図3の例では、地図の欄に記入された画像を切り出して出力している。

（5）事前に登録した画像の出力

情報センタで事前に登録した画像を出力する。図3の例では、帳票上の絵・写真の選択欄にマークすることにより、HTMLファイルには事前に登録しておいた画像（例えばカラー画像や動画）を出力する。これにより情報提供者は情報センタから提供される画像を取捨選択するだけで、簡易にカラー画像や動画を含んだ多彩かつオリジナリティの高いホームページを作成することができる。

（6）記入内容による出力制御

帳票の記入内容に応じて、文字列及び画像の出力を制御する。図3の例では、取扱い商品一覧において商品名及び価格欄に記入された文字列に合わせて、表の行数を2つに絞り込んで出力することができる。

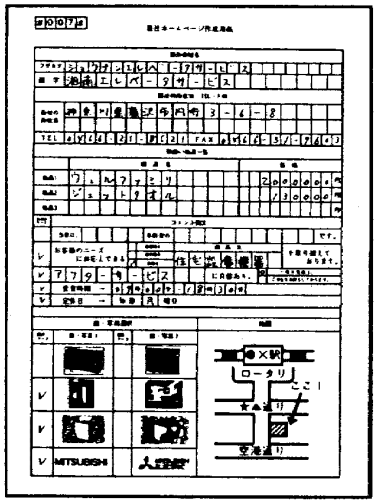
4．おわりに

文字及びマークの認識結果を元にHTMLの出力形態を変更することができるスクリプト言語を導入し、カラー画像や動画を含んだ多彩なホームページを簡易に作成できることを確認した。今後、さらに表現力の高いホームページ作成方法を検討する。

参考文献

[1] 宮本,他:"FAX文字認識によるホームページ作成法の検討",画電学会 MITC 第7回研究会 MT7-S2-3(1996)

[2] 小林,他："情報発信のためのFAX-OCR適用法",画電学会MITC第7回研究会MT7-S3-1(1996)



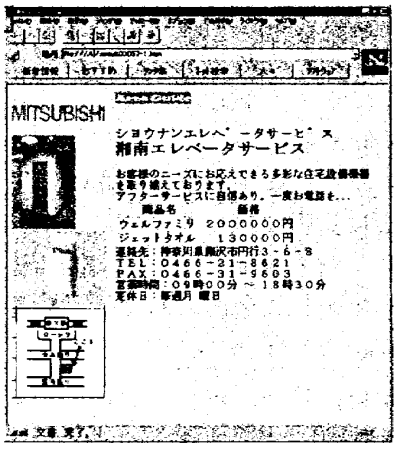
(a) 入力画像の例

```
#IF<13>
当社は、<#14#>年創業の<#15#>です。<BR>
#ENDIF
#IF<16>
お客様のニーズにお応えできる<<17|18|20>><#19#>
を取り揃えております。<BR>
#ENDIF
#IF<21>
<#22#>に自信あり。<<23|24>><BR>
#ENDIF

<TABLE BORDER>
<TR><TH>商品名<TH>価格
#IF<7>
<TR><TD><#7#>
<TD ALIGN="right"><#8#>円
#ENDIF
#IF<9>
<TR><TD><#9#>
<TD ALIGN="right"><#10#>円
#ENDIF
#IF<11>
<TR><TD><#11#>
<TD ALIGN="right"><#12#>円
#ENDIF
</TABLE>
```

(b) 出力フォーマットの例（一部）

```
お客様のニーズにお応えできる多彩な住宅設備機器を取
り揃えております。<BR>
アフターサービスに自信あり。一度お電話を..<BR>

<TABLE BORDER>
<TR><TH>商品名<TH>価格
<TR><TD>ウェルファミリ
<TD ALIGN="right">２０００００００円
<TR><TD>ジェットタオル
<TD ALIGN="right">１３０００００円
</TABLE>
```

(c) HTMLファイルの例（一部）



(d) 画面出力の例

図3　HTML自動変換の例