

印刷気象データ光学読取装置*

山元龍三郎**・林 周行***・岩嶋樹也**

要旨

印刷された気象・気候データの磁気テープ化を光学的に行う装置を試作した。紙質、活字の大きさ・形、書式などが任意であっても使用可能であって、正読率90%以上のものを作った。スキャナグラフで印刷紙面上の各点での白黒の情報を得、それを中型電算機で処理する。文字グループの位置を決定し、次に輪郭特徴抽出法により、左右から見た輪郭の特徴により、文字認識を行う。対象を数字に限定したが、処理後のチェック・修正は不可欠である。カードパンチなどの従来の手順との比較結果も述べる。

1. 緒言

気象学における観測データの重要性は周知の事である。気象現象は空間的・時間的に変動が著しいので、その研究に用いられるデータは膨大な量になる場合が多く、それ故、観測データの解析が、電子計算機を用いて行われている。最近では、ごく少数の例外を除くと、ほとんどの気象観測データは、観測直後または受信直後に磁気テープに収納されて、カードパンチ等の手順を経る事なく、電算機処理に供せるようになって来た。

気候変動の研究では、電算機時代より前に印刷刊行された観測データを用いる事が多く、長期にわたる広域の気候変動の研究に供せられる印刷データの量も膨大である。一方、それらのデータ解析に際して、より高度の統計処理を行う場合も増え、計算量も著しく増加している。それ故、過去の気象データを直接電算機処理に供せるように、磁気テープに書き込む事が要望され、各国で努力がはらわれて来た (Jenne, 1975: Panel on Data, 1980) 過去のデータを、逐一カードパンチして磁気テープに書き込む作業をしている国もある。これに対して、最近の情報科学技術などの進歩に伴って普及しつつあ

る光学文字認識装置 (Optical Character Recognition, OCR) の技術を適用する事が考えられ、その試みがこの研究である。

実用レベルにある OCR の読取の正読率は、印刷英数字については 99.9% 以上とされている (橋本, 1982)。これらの実用に供されている OCR の大部分は、入力文字の字体・大きさや書式等を制限している。これらの制限をつける事なく印刷データを光学的に自動認識する試みも報告されている (Andrewsky, 1969)。しかし、気象庁刊行の Aerological Data of Japan のように、活字の大きさが、1 mm × 0.5 mm 程度またはそれ以下であり、必ずしも良質でない用紙に、任意の書式で印刷されたデータに適用可能な市販の OCR は、筆者の知る限りでは、見当たらない。そのような印刷データの光学認識装置の製作を、この研究で試みた。研究発足時には、数字の他に、英字・特殊記号まで認識する事を目標としたが、当面、数字と 2、3 の特殊記号の認識に限る事とした。

2. 既存の光学文字認識装置の概要

1928年にオーストリアの G. Tauschek が印刷数字の光学文字認識装置 (OCR) を製作したが、OCR の誕生とされている (橋本, 1982)。数字の字体をあらかじめ切り取ったテンプレートのような面を通して、白紙に黒く印刷された数字からの反射光を受ける。入力数字とテンプレートの数字とが一致しない時には、印字以外の白紙の部分からの反射光が受けられるが、一致した時

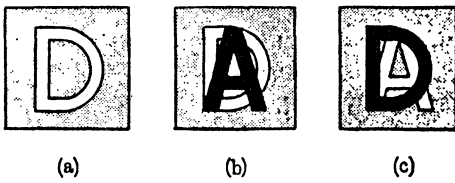
* An optical character recognition for printed data of meteorological observations.

** Ryuzaburo Yamamoto and Tatsuya Iwashima, 京都大学理学部気候変動実験施設。

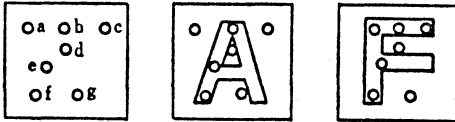
*** Kaneyuki Hayashi, 株式会社コムシステム技術部。

—1982年8月5日受領—

—1982年11月4日受理—



第1図 テンプレートマッチング (橋本, 1982).



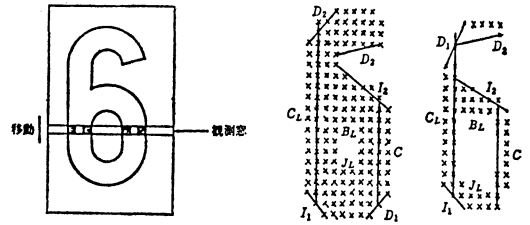
第2図 定点マッチング (橋本, 1982).

にはテンプレートを通る反射光は零となるので、数字の光学的認識が出来る。これが、Tauschek の OCR 原理である。

1950年代のデジタル電子計算機の実用化に伴って、近代の OCR が誕生する事となった。米国の Farrington 社の OCR が 1955 年に発表されて以来、英国・米国などの各社で、種々の装置が製作され始めた。1957 年には、手書き文字認識の試みがベル電話研究所で行われた。郵便番号読取用手書き数字 OCR を東京芝浦電気(株)が製作したのが1966年である。1973年には、印刷漢字認識実験が始められた。1970年代後半では、印刷英数字、手書き数字、仮名の OCR は、ほとんど完成し、1970年代末からは、印刷漢字 OCR などの実用化の研究が活発になって来た。

以下においては、印刷文字の認識について概説し、手書き文字については省く事とする。まず、印刷文字の白黒のパターンを光电変換して電気的信号に変える。この信号から文字を認識するのに、いくつかの方法が考案されている (橋本, 1982)。第1はパターンマッチング法と呼ばれるものである。これには入力パターンとあらかじめ用意した標準の型との比較によって識別するテンプレートマッチング (第1図)、あらかじめ位置を定めた若干個の覗き窓から入力パターンを見て、各窓の黒・白の組合せから文字を認識する定点マッチング (第2図)、文字パターンを2次元配列のメッシュに分解し、入力パターンと標準パターンの行列の比較により、文字認識を行うマトリックスマッチングがある。これらの手法では、字の形や大きさが任意であるような場合の認識は容易ではない。

第2の方法は、ストロークアナリシス法と呼ばれるも



第3図 外部特徴抽出 (橋本, 1982).

- (a) 1 行の観測窓によるパターン左右端の検出。
- (b) 外部形状の接線近似。

A B C D E F G H I J K L M
 N O P Q R S T U V W X Y Z
 0 1 2 3 4 5 6 7 8 9

OCR-A フォント

0 1 2 3 4 5 6 7 8 9
 A B C D E F G H I J K L M
 N O P Q R S T U V W X Y Z
 a b c d e f g h i j k l m
 n o p q r s t u v w x y z

OCR-B フォント

第4図 OCR-フォント。

ので、文字パターン線の構成に着目したものである。文字枠の内に、水平な線分窓を上端・下端・中央に設定し、また、鉛直線分を左端と右端に設ける。これらの線分の窓から入力パターンを見た時の白・黒の組合せによって、文字を判別するのである。この方法でも、認識出来る字の形に制限がある。

第3は、輪郭特徴抽出法であって、文字の輪郭を追跡する手法と、外部の特徴を抽出する手法とがある (第3図)。後者の方法を、この研究では採用しているので、詳細は後述する。

市販の印字用 OCR の多くは、OCR フォント (font) と称せられる認識の容易な活字を対象としている (第4図)。その文字の寸法は、高さが2.4~3.8 mm、幅が1.4~2.0 mm である。OCR フォント以外の活字をも

STANDARD PRESSURE LEVELS

CM	Station Sendai																			
	1000 mb																			
	(°C)	U(%)	d	w(m/s)	sp(ppm)	T	U	d	w	sp	T	U	d	w						
0	2.0	73	300	2.7	111	2.6	74	271	3	969	5.5	61	216	10	1436	4.6	39	237	12	
1	0.8	59	40	2.3	105	0.1	62	33	3	939	-7.0	80	297	7	1382	-10.4	90	305	11	
	-1.4	67	340	*0.8	214	-2.4	66	309	3	1041	-9.2	72	305	9	1479	-12.8	78	303	15	
2	-1.6	71	320	1.2	226	-0.7	63	242	2	1073	0.3	*40	229	6	1529	-2.0	34	248	8	
	6.6	54	240	1.7	117	7.2	57	226	5	986	6.9	68	240	8	1452	2.6	80	259	12	
3	1.0	51	340	3.7	212	-0.2	55	344	5	1043	-7.5	73	308	4	1486	-11.7	84	310	9	
	-3.7	71	360	2.2	*276	-4.6	70	11	4	1105	-5.7	46	43	2	1553	-4.5	27	75	*4	
4	0.7	72	340	2.0	256	0.4	66	327	4	1095	-4.5	77	303	11	1543	-8.0	80	298	14	
	-0.2	79	10	1.3	272	-0.4	78	2	4	*1112	-2.5	56	283	8	*1565	-1.9	34	304	6	
5	0.2	79	350	2.5	251	1.8	71	16	4	1095	-1.3	54	319	6	1550	-1.3	*26	339	5	
	3.0	81	340	3.3	204	3.0	91	20	5	1060	1.4	100	136	12	1518	-1.3	100	147	12	
6	7.5	95	360	4.0	63	*7.5	95	1	4	929	4.6	97	*86	12	1392	1.4	100	86	12	
	6.9	79	350	5.5	64	6.8	80	350	6	927	4.0	86	10	8	1389	0.9	93	41	9	
7	1.8	79	330	3.2	181	1.8	75	356	6	1028	-2.3	77	357	*1	1479	-3.4	53	349	*4	
	3.1	74	340	3.8	155	2.7	76	345	4	1009	0.4	88	138	6	1464	-3.0	90	176	7	
8	7.0	59	340	6.0	*56	6.9	59	339	6	*913	0.7	69	323	14	*1370	-1.5	59	334	22	
	1.6	57	360	5.5	170	0.8	58	346	8	1007	-5.7	66	304	10	1451	-10.1	74	307	15	
9	0.2	64	110	1.7	237	-0.5	64	236	4	1072	-3.9	56	275	7	1524	-4.7	39	327	*4	
	1.7	73	150	*0.8	186	3.0	71	-	*0	1041	1.0	83	274	14	1499	-1.8	90	281	18	
10	4.5	83	20	1.5	130	4.5	84	21	2	985	1.0	90	207	4	1443	-1.2	91	241	7	
	1.8	93	320	1.8	148	1.6	94	316	1	995	-0.6	100	148	7	1450	-3.3	100	207	7	
11	0.9	72	270	3.3	92	0.6	73	271	4	929	-5.4	73	299	19	1376	-7.9	74	300	23	
	2.0	73	30	1.0	191	1.5	74	44	1	1035	-3.3	83	276	12	1486	-5.5	87	273	19	
12	-2.0	63	20	2.5	240	-3.0	71	4	4	1074	-2.4	87	219	5	1528	-3.0	90	237	8	
	1.4	88	350	2.8	86	1.3	88	351	3	940	*7.9	84	250	8	1410	*5.4	93	239	9	
13	3.3	55	350	3.5	102	2.9	55	352	4	945	-4.5	64	300	10	1392	-8.4	73	300	14	
	0.4	64	20	2.0	162	-0.6	65	21	2	996	-5.6	71	272	5	1441	-8.8	74	282	10	
14	0.1	58	40	*0.8	162	-0.1	54	332	3	998	-5.3	51	297	10	1444	-8.3	56	287	16	
	1.1	51	310	2.2	215	-0.3	54	303	6	1046	-7.8	64	270	7	1488	-11.1	71	256	8	
15	6.1	65	270	*6.2	76	5.8	66	271	7	930	0.6	72	287	*24	1387	-2.3	64	288	*24	
	1.9	*49	290	5.8	185	0.9	*51	286	*12	1021	-6.0	52	307	16	1467	-7.5	42	313	21	
16	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	
	1.9	69	31	2.8	166	1.6	70	4.2	4	1011	-1.8	72	9.1	7	1464	-4.2	70	11.8	8	
17	W 54 E 9 S 6 N 90												200	30	46	90	266	29	49	131
	W 20.5 E 10.5 S 3.5 N 27.5												25	6	9.5	21.5	26	5	10	21
18	W 1.5 S -2.7 * 331 R 3.1												5.5	-1.4	284	5.7	7.6	-2.6	289	8.0
	-1.6	71	350	2.3	164	-2.2	71	4	3	996	-6.8	77	274	7	1441	-8.8	78	278	12	
19	0.7	51	280	6.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

第5図 気象庁刊行の Aerological Data of Japan の印刷の一例。スケールは mm.

対象にする装置の場合でも、活字の大きさは、同程度のものである。したがって、第5図に物差しと共に示した Aerological Data of Japan のように、1mm 程度の大きさの任意の形の活字を認識する市販の OCR は見当たらない。これは、実用レベルの OCR では正読率が 99.9% 程度の値をもつ必要があり、そのような高い正読率を、良質ではない紙に印刷された任意の活字・書式のものに対して得る事が、技術的に困難であるからであろう。それ故、この研究では、Aerological Data of Japan など過去の印刷気象データについて、正読率が90%以上になるような OCR を製作する事を目標とした。

3. 試作装置とその操作の概要

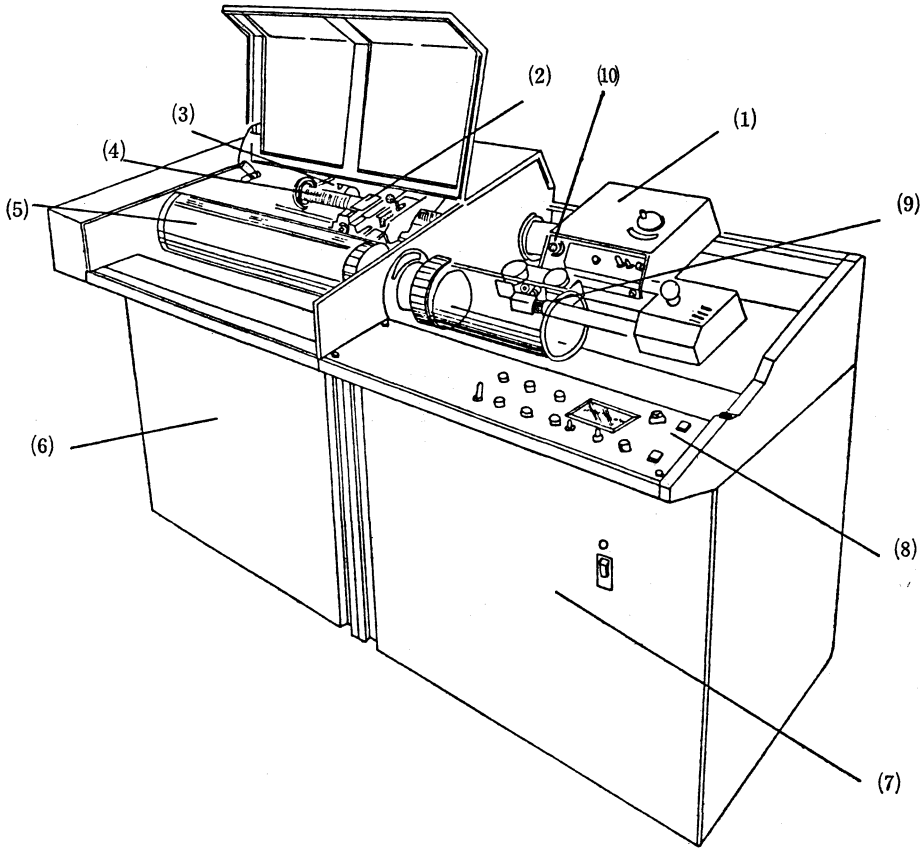
印字を電気的信号に変換するために、この試作装置では、大日本スクリーン製造(株)製のスキヤナグラフ SG-100 K を用いる事とし、その出力を TEAC (株)製 TFC-161 の磁気テープ装置に記録する。このスキヤナ

第1表 スキヤナグラフ SG-1000 K の主な仕様。

入力印刷データの有効面積	35 × 25 cm ²
ドラム回転速度	3 RPS
横送り方向分解線数	500本/インチ
サンプリングレート	約 40 μsec
入力側サンプリングアパーチャ	50μm × 50μm
ADコンバータ分解能	256分割 (8 bit)
走査速度	線170秒/インチ

グラフの概略図を第6図に、その主な指標を第1表に示す。

このハードウェアの具体的操作は、第6図で示されている原稿シリンダーの外側に、入力印字データを貼付する。その際、印刷データを、第7図の左端に示したバー



第6図 スキャナグラフの概略図。

- | | | |
|------------|----------------|-----------------|
| (1) 走査ヘッド部 | (5) 記録シリンダー部 | (8) コントロールボックス部 |
| (2) 記録ヘッド部 | (6) 原稿シリンダー収納部 | (9) 原稿シリンダー部 |
| (3) 駆動モータ部 | (7) 電源及び駆動回路部 | (10) のぞき窓 |
| (4) 送りネジ部 | | |

インデックスをあらかじめ画いた透明シートで、おおって、原稿シリンダーにはり付ける。このパーインデックスは、印刷行の印字の中央に位置するようにする。印刷データがなるべく傾かないように、原稿シリンダーを回転させながら、覗き窓から野線を監視する。

印刷データの紙面上で、読取範囲を定め、その値を設定した後、スキャナグラフによる走査を開始する。第1表に示すように、原稿シリンダーはその軸の周りを1秒間に3回転するが、1回転する毎に、 $50\mu\text{m}$ だけ走査ヘッドが軸方向にずれる。原稿シリンダーの全円周の約1/4の間では、受光されないで、残りの有効画面範囲に入力印字データを貼付する必要がある。円周方向の分解能も $50\mu\text{m}$ であって、その方向に10インチだけ走査する場合、一周分のスキャナデータは5,000バイトとな

る。軸方向の走査範囲を10インチとした場合、原稿シリンダーは5,000回転して、この範囲を走査するので、このスキャナグラフからのデータは正味 5×10^3 バイト $\times 5 \times 10^3 = 25 \times 10^6$ バイトとなる。これは、1,600 bpiの密度で磁気テープに書き込む時、2,400フィート磁気テープの1巻におさまる。10インチ分の走査には、約28分の時間を要する。このように得られたスキャナデータは、オフラインで次章に述べるように、電子計算機システムによって数字認識に供される。認識の結果は、ディスプレイ装置上で、チェックおよび修正を行う。最終結果は、別の磁気テープに書き込み、気象・気候の本来の研究に供する。

4. スキャナデータからの数字認識

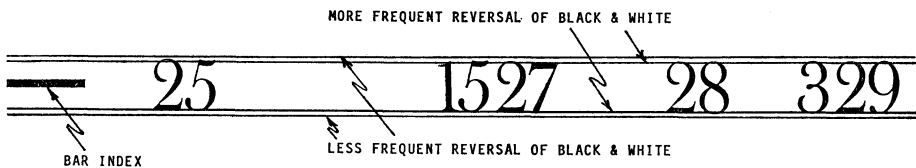
スキャナグラフにより、印刷データの紙面の白黒の情

STANDARD PRESSURE LEVELS

Date	Station Tatteto																																							
	300 mb			250 mb			200 mb			175 mb			150 mb			125 mb																								
	sp (µm)	T (°C)	U (k)	U (m/s)	sp	T	U	d	e	sp	T	U	d	e	sp	T	U	d	e	sp	T	U	d	e	sp	T	U	d	e											
1	9255	-28.1	257	70	10491	-45.2	254	80	11961	-51.1	257	80	12824	-54.2	257	72	13602	-57.6	261	69	14940	-63.9	259	56	14940	-63.9	259	56	14756	-59.4	245	60	14706	-52.7	270	37	14822	-57.0	261	45

第7図 バーインデックスとその配置法。

BAR INDEX



第8図 行の上・下端の決定。

報が縦・横ともに50μmの分解能で得られて、それらが磁気テープに収納される事を前章で述べた。このスキヤナデータから、オフラインで、電子計算機システムM150-F(富士通(株)製)により、数字認識を行う。その際、前もって、1行の長さを若干の余裕をもって指定しておく。また、1文字分の枠の横幅および高さを0.1mm単位で指定する。数字認識の処理は、以下に述べたような段階に分けられるが、これらは、1つのマクロコマンドのキーインにより実行出来る。

4.1. スキヤナデータの入力

数字の認識は、1行毎に行うが、文字枠の高さの約2倍の高さ範囲のスキヤナデータを読み込んで、以下の処理を行なう。その行の文字認識が完了すれば、既に読み込んでいるスキヤナデータの上半分の代わりに、新しく、すぐ下の行の文字枠の高さ範囲のスキヤナデータを加える。これらのスキヤナデータにより、次の行の数字認識を行う。このように順次、各行を処理する。

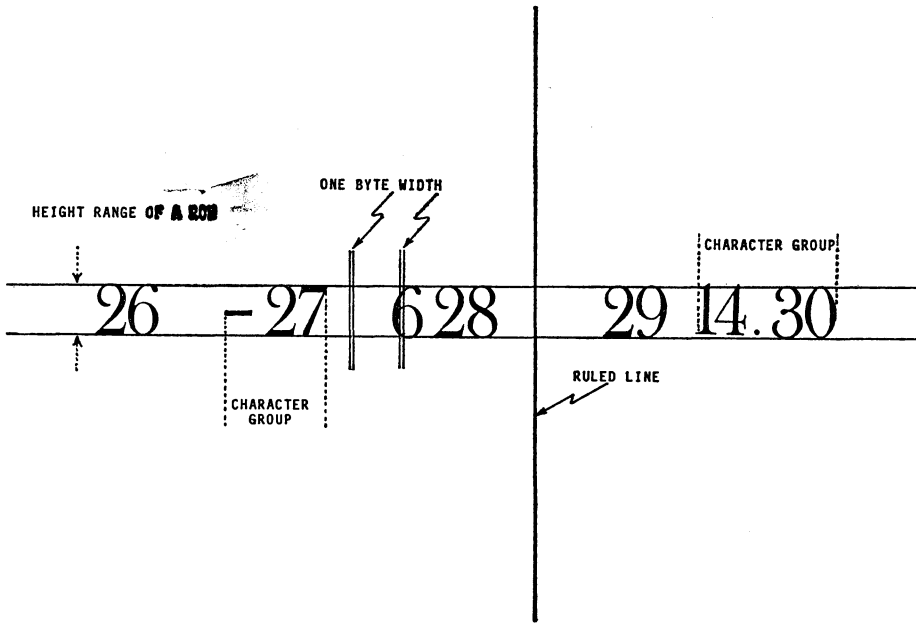
4.2. 文字の上・下端の決定

第7図に示したバーインデックスは、十分な長さとな

きをもつので、容易に識別出来る。このバーインデックスは、印字行のほぼ中央に位置するように調整されているが、正確にそのようになっていない。それ故、文字行の上下端の決定をまず行う。バーインデックスを中心にして、文字枠の高さを11.672に0.6mmを加えた高さ範囲を、横方向に走査する。そして、白・黒の反転回数の変化によって文字の上端・下端を決定する(第8図)。すなわち、白黒の反転回数が、文字行の上・下端をよぎって、唐突に変化する事に着目して、上・下端の位置を決定する。1つの行に対して、このような処理を左から右へ、1行の1/4の長さ毎に行う。印刷データをスキヤナグラフの回転シリンダーに取り付ける際に、少し位傾いていても、このような処理により、後段の処理においても支障が起らないようにする事が出来る。

4.3. 各行における文字解読の基本方針

行の上・下端を確認した後、その行の文字の解読に進む。その際、1文字ずつ解読しないで、文字グループ、すなわち、第9図の例で示すと、26・-27・628・29および14.30をまとめて検出する。そして、それぞれのグ



第9図 文字グループの決定.

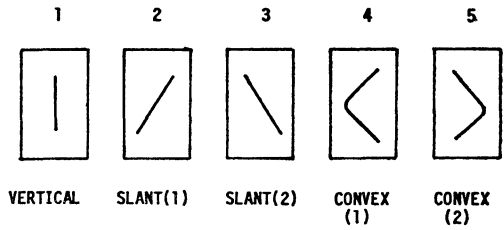
ループの右端の位置が、パーインデックスからどれだけ離れているかによって、そのグループの位置を決定する事とした。左から1文字ずつ順次解読する方法では、行の右端近くで、大きく桁ずれを起こす心配がある。また、'123' というグループを '12+3' と判断したり、'13' と認識したりする懸念がある。これらを防ぐために、上述のように、文字グループの検出をまず行う事とした。

4.4. 文字グループの検出

1つの行の上・下端は既に決定出来ているので、その高さ範囲内で、横幅1バイト(50 μm)毎に順次、白黒の点の数を調べる(第9図)。1行の高さのほとんど全体にわたって黒の場合は、データとして意味のない野線と判断して、以後の処理を行わない。上・下方向に、ほとんど全て白であるようなものが横方向にある定数以上連続している場合には、空白とみなし、文字グループの区切りであると判断する。白の連続が、ある定数以下の場合には、同じグループの中の文字の切れ目だとする(第9図)。上・下方向に、白と黒とが混在している場合は、そこに文字があるものと判断し、以後の処理の対象とする。

4.5. 個々の文字の解読

第2章で述べたように文字の解読手法には種々のものがあるが、ここでは輪郭特徴抽出法を取り上げる事とす



(a)



4 4 4 3 4 2 -----> 3

(b)

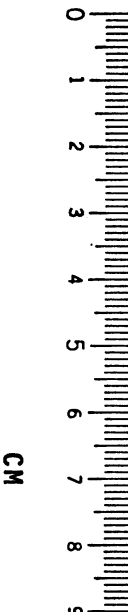
第10図 外郭特徴抽出による文字解読.

(a) 5種類の特徴的型.

(b) 数字3の場合の特徴的型の組合せ.

る。これは、1959年に W. Doyle の提案したものであって (Selfridge and Neisser, 1960), 数字のように文字の種類が少ない場合には有用であるとされている (坂井

SURFACE DATA JANUARY 1980



LATITUDE	LONGITUDE	ELEVATION METERS	NUMBER OF DAYS OF OBSNS.	PRESSURE		TEMPERATURE		VAPOR PRESSURE		PRECIPITATION			SUN- SHINE	
				MEAN STATION MB	MEAN SEA LEVEL MB	MEAN °C	DEPARTURE °C	MEAN MB	DEPARTURE MB	NO. OF DAYS i MM.	TOTAL MM	DEPARTURE MM	QUINTILE	PERCENTAGE OF LONG-TERM AVERAGE %
22 12 N	113 32 E	59	31	1011.7	1018.7	15.4	+ 0.3	13.6	+0.6	2	5	- 21	2	117
37 45 N	128 54 E	27	31	1016.1	1019.6	2.6	+ 3.6	3.7	+0.6	5	58	+ 21	4	85
37 29 N	126 38 E	70	31	1013.6	1022.7	- 3.4	+ 0.6	3.3	+0.1	6	30	+ 14	4	80
35 06 N	129 02 E	71	31	1011.5	1020.3	2.4	+ 0.6	4.2	+0.5	4	18	- 7	4	83
34 47 N	126 23 E	56	31	1015.5	1022.6	1.8	+ 0.8	5.4	+0.7	6	52	+ 15	4	85
45 25 N	141 41 E	11	31	1008.7	1010.1	- 5.3	+ 0.5	3.1	0.0	28	146	+ 42	5	65
43 46 N	142 22 E	116	31	996.3	1011.2	- 7.2	+ 1.3	3.1	+0.3	23	76	- 5	3	119
44 01 N	144 17 E	39	31	1004.0	1009.5	- 5.6	+ 1.0	3.1	+0.2	15	77	+ 16	4	83
43 03 N	141 20 E	19	31	1008.7	1011.1	- 3.9	+ 1.2	3.2	-0.1	19	97	- 21	2	111
43 20 N	145 35 E	26	31	1004.7	1008.1	- 3.5	+ 1.3	3.5	+0.2	17	105	+ 57	5	80
42 10 N	142 47 E	34	31	1005.6	1009.9	- 2.6	+ 0.7	3.5	+0.1	12	64	+ 17	4	103
39 43 N	140 06 E	10	31	1012.5	1013.7	- 0.1	+ 0.6	4.7	+0.2	22	173	+ 38	5	131
39 39 N	141 58 E	47	31	1006.0	1011.8	0.8	+ 0.9	3.9	+0.2	2	55	- 8	3	97
38 16 N	140 54 E	40	31	1008.5	1013.5	1.4	+ 0.8	4.7	+0.2	8	28	- 14	2	99
37 23 N	136 54 E	6	31	1014.9	1015.7	2.8	+ 0.3	6.2	+0.5	26	352	+ 71	4	88
37 55 N	139 03 E	6	31	1013.9	1014.7	2.6	+ 0.8	5.7	+0.3	22	225	+ 19	3	72
36 33 N	136 39 E	28	31	1012.6	1016.1	3.1	+ 0.5	6.0	+0.1	23	277	- 50	3	79
36 15 N	137 58 E	611	31	941.4	1015.5	0.1	+ 1.3	4.2	+0.3	8	36	0	3	82
36 24 N	139 04 E	113	31	1000.1	1014.1	3.0	+ 0.4	4.4	+0.3	5	29	+ 4	3	92
35 10 N	136 58 E	56	31	1009.4	1016.3	4.1	+ 0.9	5.5	0.0	9	78	+ 26	4	76

第11図 Monthly Climatic Data for the World 中の surface data の一部。

・長尾, 1971).

文字グループの位置を決定した後、そのグループ内の個々の文字を、次の手順により解読する。

(i) あらかじめ与えた文字枠内で、左および右から見た時の輪郭を求める(第10図)。輪郭が枠の高さの1/2以下の場合、数字ではなくて、小数点などの特殊記号と判断する。その高さとの比などから、マイナス記号または小数点を区別し、区別の出来ない場合には「*」として出力する。

輪郭が枠の高さの1/2以上の場合には、数字であると判断して、以下の処理を行う。

(ii) 第10図(b)に示したように、枠を上下方向に3等分した区画に分割する。これによって、左右から見た輪郭が6個に分けられる。このように分けられた輪郭の特徴を、第10図(a)に示した5個の型(垂直, 右上り, 左上り, 左に凸, 右に凸)のいずれかに対応させる。

(iii) 6個の分割された輪郭のそれぞれに対して、5個の型から1個ずつを選ぶ順列となるので、合計 $5^6=15,625$ 通りの場合が起こり得る。実際に数字となる場合は、約900通りであって、これらを、マッチングテーブルとして、記憶装置上にあらかじめ保存しておく。6個の輪郭特徴の型の順列を、このマッチングテーブルと照合して、一致する場合をさがす。一致した時、その数字

を解答とする。もし、一致するものが見当たらない場合は、解読不能として「?」を出す。

(iv) 上述の、輪郭の外側の特徴的型のみでは、識別しにくい場合がある。印刷が不鮮明な時は勿論だが、活字の形によっては、例えば、「8」と「6」のように、まぎらわしい場合が時に起こる。このような場合に対して、マッチングテーブル上で、複数個の数字を与えておいて、次の手順のいずれか、または両方を用いて、判断する。

(a) 閉曲線で囲まれた大部分(穴)の有無およびその数を調べる(例えば、「9」「8」「7」)。

(b) 3等分した区画の最上部について、左右から見た時に、影となる部分の面積の大小の比較(例えば「5」と「3」については、「3」では影が広い)。

電子計算機システム M150-F による上述の処理に要する時間は、Aerological Data of Japan の Standard Pressure Level の 1/2 頁分(1行の文字グループが31, 1グループ内の文字の数が1ないし5, 空白行を含む行の数が37)に対して約25分である。

以上の処理によって得られた解読結果は、計算機システムに記憶される。Aerological Data of Japan では、月間の極値を示す特殊記号が用いられたり、印刷が不鮮明であったりするので、99%またはそれ以上の正読率が

第2表 Monthly Climatic Data for the World の地上データに対するテスト結果の1例.

入 力		出 力			
入力文字	入力文字数	正 読 率 (%)	認識不能率 (%)	誤 出 力	
				誤出力文字	誤出力率
0	336	82.7	6.5	1 3 7 8	2.7 3.0 4.8 0.3
1	550	95.3	2.0	7	2.7
2	188	94.7	2.7	7	2.7
3	327	97.9	2.1	—	0.0
4	172	99.4	0.0	0	0.6
5	172	96.5	3.5	—	0.0
6	132	93.9	6.1	—	0.0
7	132	99.2	0.8	—	0.0
8	139	97.8	0.7	0	1.4
9	124	95.2	1.6	3 7	0.8 2.4
平 均	2272 合 計	94.5	2.8		2.8

期待出来ないので、電算機による解読結果をディスプレイの画面に出して、チェック・修正をする必要がある。チェック・修正の終わった結果は別の磁気テープに順次、書き込んで、このOCRによる作業が終了する。

5. 試作 OCR のテスト

上に述べた試作OCRを、電子計算機システムM150-Fを用いて、テストした。まず、米国のNOAA, National Climatic Center 刊行の Monthly Climatic Data for the World に掲載されている surface data を取り上げた。その一部を、文字の大きさを示すために、物差しを添えて再録したのが、第11図である。この印刷データを、試験的に5回自動認識した結果、正読率は98.3~92.1%の範囲にはいり、その平均は95.1%であった。そのうちの一例について、結果の詳細を示したのが第2表である。0から9までの入力文字のそれぞれについて、入力数、正読率を与えてある。さらに、認識不能で「?」が出力された頻度、および入力とは異なる数字を誤出力した時の頻度を与えている。0および6の認識不能、および0の入力に対する誤出力が目目される。なお、ここでは英字は対象としていないので、正読率の計算では考慮していない。正読率のばらつきは、パーインデックスの配置・印刷紙の紙質・印字のかすれ程度、

しみの有無などによると考えられる。

次に Aerological Data of Japan (その一例が第7図に示してある) についても、同様なテストを5回行った。その結果、平均の正読率は91.8% (98.9~88.3%) であった。そのうちの一例について、結果の詳細を第3表に示す。0および8の認識不能が4%に達している。また、5および8の誤出力が2%を越えている。

活字が大きく、印刷の鮮明なデータの正読率が好成績である事は当然予期された通りである。しかし、実用レベルのOCRに比べると、この研究の試作装置で過去の印刷気象データを読み取る場合の正読率はかなり劣るので、チェック・修正は不可欠である。

印刷気象データの処理に要する時間を、カードパンチの場合と比較するために、次のような概算をした。Aerological Data of Japan (Standard Pressure Level) の1/2頁には、約3,000個の数字があり、空白・改行などを含めると、カードパンチの際のストローク数は約4000回と算定される。専門のキーパンチャーは1時間当たり7,000ストロークする能力があるとすると、上記のデータのパンチに約34分を要することとなる。Aerological Data of Japan のように、活字の大きさが小さい時には、あらかじめ、コーディングシートに書き写す必要があり、このために概算30分を要すると思われる。さら

第3表 Aerological Data of Japan に対するテスト結果の一例

入 力		出 力			
入力文字	入力文字数	正 読 率 (%)	認識不能率 (%)	誤 出 力	
				誤出力文字	誤出力率 (%)
0	948	95.0	4.0	3 7 6	0.6 0.2 0.1
1	1680	97.6	0.8	0	1.6
2	1047	98.9	1.1	—	0.0
3	915	99.2	0.4	7 5	0.2 0.1
4	621	97.3	1.1	0 6	1.0 0.6
5	597	95.6	1.8	9 7 6	2.0 0.3 0.2
6	549	98.2	1.5	0 5	0.2 0.2
7	702	99.0	1.0	—	0.0
8	642	91.4	4.0	3 0 6 9 2 7	2.2 0.8 0.6 0.5 0.3 0.2
9	747	99.3	0.4	3 7	0.1 0.1
平均	8448 合計	97.3	1.5		1.1

に、パンチしたカードのチェックに約30分を要するものと仮定すると、Aerological Data of Japan の1/2 頁のカードパンチに、合計約95分を要する事になる。

一方、この研究の試作装置については、Aerological Data of Japan の1/2 頁は約3インチの高さなので、スキャナグラフによってスキャナデータを磁気テープに書き込むのに、約8分を要する。これを計算機システム M150-F によって解読処理をするのに、約25分を要する。その処理結果のチェック・修正には、正読率が劣るので、45分の時間を要するものとする、合計78分の時間が必要となる。そのうち、人手を要するのは、チェック・修正とスキャナグラフの操作・スキャナデータの磁気テープのマウントなど約50分間と見積られる。

6. 結 語

過去の印刷気象データの磁気テープ化のために、光学文字読取装置(OCR)を試作した。活字の形・大きさや

印刷紙の質などについて、十分に吟味した既存の OCR の99.9%以上の正読率に比べると、印刷数字に限った試作装置のテストでは、かなり劣った正読率しか得られなかった。活字が比較的大きい場合には、約95%、小さい活字に対しては約92%であるので、このままでは、解読結果のチェック・修正が不可欠である。

しかし、カードパンチによる磁気テープ化の場合と比較すると、特殊技術を要しない事および所要時間が若干少なくてすむなどの利点がある。ここで用いた計算機よりも演算速度の速い計算機を利用すれば、所要時間は短くなる。

今後、正読率を向上させるための改良を試みる必要がある。特に、解読処理において判断不能の場合には疑問答が出力されるので、チェックの時修正が容易である。しかし、全く別の数字だと誤判断する場合もまれに起こる。このような障害を少なくするように、文字解読処理法の改良が、英字をも解読し得るように拡張する事と共

に、必要であろう。

そのために、マッチングテーブルの改良充実をする必要がある。事実、マッチングテーブルを若干充実してテストした結果、正読率が向上した。また、第4章の(iv)で述べた認識の第2段階をさらに増やす事により、正読率を高める事が出来よう。このような改善は、比較的容易に実施し得るものであるが、印刷のかすれなどは、不可避なので、正読率が99.9%以上にする事は不可能と考えられる。したがって、人手によるチェック・修正は必須であろう。一方、第4章で述べた方法とは全く異なる文字認識法によるクロスチェックを同時に行えば、正読率が著しく向上するものと考えられる。これは、今後の研究課題である。

この研究の実施に対して、トーアエレクトロニクス(株)および日本スクリーン製造(株)の多大の協力があつた事を付記して、謝意を表わしたい。

文 献

- Andreewsky, E., 1969: Research on pattern recognition in France, Methodologies of Pattern Recognition, Ed by S. Watanabe, Academic Press, New York.
- 橋本新一郎, 1982: 文字認識論, オーム社, p.258.
- Jenne, R.L., 1975: Data sets for meteorological research, Boulder, NCAR PB-246, 564, 194 p. NCAR,
- Panel on Data, 1980: Report of the panel on data, Tech. Conf. Climate Asia and Western Pacific, Proc. the Conf. Guangzhou, China, 15-20 Dec. 1980.
- 坂井利之・長尾 真, 1971: 文字・図形の認識機械, 共立出版(株), 172p.
- Selfridge, O.G. and U. Neisser, 1960: Pattern Recognition by Machine, Scientific American, 203, 60-68.

第22期第2回常任理事会議事録

日 時 昭和57年11月15日 9.45~12.00

場 所 気象庁観測部会議室

出席者 岸保, 松本, 荒井, 内田, 河村, 嶋村, 杉村,
竹内, 田宮, 二宮, 増田, 村山

議 題

1. 昭和58年度予算(案)について

前回の理事会で各理事から特に意見が出ていないので、原案にそつて更に検討することになった。

2. 昭和58年度春季大会について

来年5月18~20日に筑波の研究交流センターで開催することが報告された。シンポジウムは大気境界層を予定している。

3. 事務局職員の採用について

推薦された候補者に岸保理事長がそれぞれ面接を

行い、理事長を含めた選考委員会で最終決定をすることになった。

なお、事務局職員の任期について討論し、現行の8年を4年にするように内規を改めることになった。

4. 韓国文化センターへの学会誌の寄贈について

「天気」「気象集誌」を寄贈することが承認された。

5. その他

(1) 長期計画委員会の報告

長期計画委員会でまとめられた報告は、予算の許す範囲内で「天気」に掲載することになった。

(2) 熱帯気象学に関する地域科学会議

米国気象学会から記念のボールをいただいた。
承認事項 藤井 享ほか12名の新入会員を承認。