Hitotsubashi Journal of Arts and Sciences 45 (2004), pp.1-9.

DEVELOPMENT OF AN OPTICAL CHARACTER READER (OCR) FOR READING PRINTED OLD CATALOGS OF ASTRONOMICAL DATA*

KOICHI NAKAJIMA

I. *Introduction*

The list of the astronomical data, e.g. a list of stars in the sky, is called "Astronomical Catalog". The most well-known example is the "Messier Catalogue", which is a list of non-stellar objects such as nebulae, star clusters, or galaxies in the sky.

In order to utilize these data for astronomical researches, it is very important that these catalogs are provided in "machine-readable form" for utilizing them on the computer. For this purpose, a world-wide project of collecting as many astronomical catalogs as possible, and providing them as machine-readable data, has been carried out among astronomers in the world.

This project is often called "catalog archiving". An institute in France, "Centre de Donnees astronomiques de Strasbourg" (CDS) in the Strasbourg Astronomical Observatory, has been one of the Centers of this project since 1972, and many institutes in the world are cooperating for this project, utilizing a contemporary tool "the Internet".

The role and works of CDS was reviewed by the author (Nakajima 1993), and the world cooperation of catalog archiving was reviewed in Nakajima (1994). A comprehensive survey of the history, meaning, methods, etc. of catalog archiving is given by Jaschek (1989).

Among many works about catalog archiving, there is a work which is steady and not conspicuous, that is, collecting old astronomical data and archiving them as machine-readable catalogs. In CDS, a part-time worker is engaging in this work. He/she often uses the OCR for reading and compiling old printed astronomical data.

Among such old data, there are catalogs which are very important but were published only in the form of printed tables. As a typical example, "The Astrographic Catalogue (AC)" which was published almost a hundred years ago, has been recently compiled into machine-readable catalogs. The AC catalog contains some 4.6 million stars' data. This difficult and laborious work has been attained by two groups of American and Russian astronomers (Urban et al. 1997, Gulyaev & Nesterov 1992).

The present paper reports a work of compiling such an old astronomical catalog. The target catalog is "MK Spectral Classification, 6th General Catalogue" by Buscombe (1984). Though it was originally compiled on a computer, the machine-readable data have been lost accidentally after publishing the computer-printed tables. Since the other tables of this series

* The author expresses his hearty thanks to all the staffs of the Astronomical Data Analysis Center, National Astronomical Observatory Japan. He also expresses his deep gratitude to Dr. Shiro Nishimura for introducing him the necessity of this work.

are already archived in the Internet database, the compilation of this catalog is awaited for long time.

In the present work, the author describes mainly the development of an OCR (Optical Character Reader) system for reading and compiling the printed table. Because the general-purpose OCR softwares available at present are all very insufficient, a special purpose OCR system is needed in which the parameters for managing characters can be adjustable finely and widely.

This work is carried out as one of the projects by the astronomical database service group in the Astronomical Data Analysis Center (ADAC) of National Astronomical Observatory Japan, and supported by it.

II. *The Catalog*

The target catalog is a part of a series of supplements for a pioneering catalog, "Catalogue of Stellar Spectra Classified in the Morgan-Keenan System" by Jaschek et al. (1964). It contains about 20000 stars' spectral type data.

First supplement, i.e. the 2nd catalog of this series, was compiled and published by Kennedy and Buscombe (1974) (later, a new version of was published by Kennedy 1983). A "Selected Catalogue" made by joining these two catalogs was provided by Jaschek (1978).

From the 3rd to 15th supplement catalogs of the series have been published by Buscombe (1977, 1980, 1982, 1984, 1988, 1990, 1991, 1992, 1994, 1995, 1998, 1999, 2001), often with a co-author B.E. Foster.

The archived catalogs accessible from Internet are provided by CDS and/or Astronomical Data Center (ADC), NASA, USA. Note that NASA ADC has withdrawn from the work of collecting and compiling new catalogs, in October 2002. Both archives are mirrored at ADAC in http://dbc.nao.ac.jp/prt3.html, as follows:

3018B : 1st catalog by Jaschek (1964)
3019 : 2nd by Kennedy (1974), superseded version
3078 : new 2nd by Kennedy (1983)
3052 : 3rd by Buscombe (1977)
3228 : 4th by Buscombe (1980)
3189 : 7th-11th by Buscombe (1988-1994)
3189A : 12th by Buscombe (1995), superseded by 3223
3223 : revised 12th by Buscombe (1995)
3206 : 13th by Buscombe (1998)
3222 : 14th by Buscombe (1999)
3225 : 15th by Buscombe (2001)

Note that NASA ADC's archive of 3052 is erroneously superseded, and that of 3189A is erroneously noted the destination of superseding one (3206, instead of correct 3223).

A machine-readable data of 5th supplement is stored in ADAC database, but not yet opened. The 6th one is the present target catalog whose machine-readable data have been lost accidentally (Buscombe, private communication).

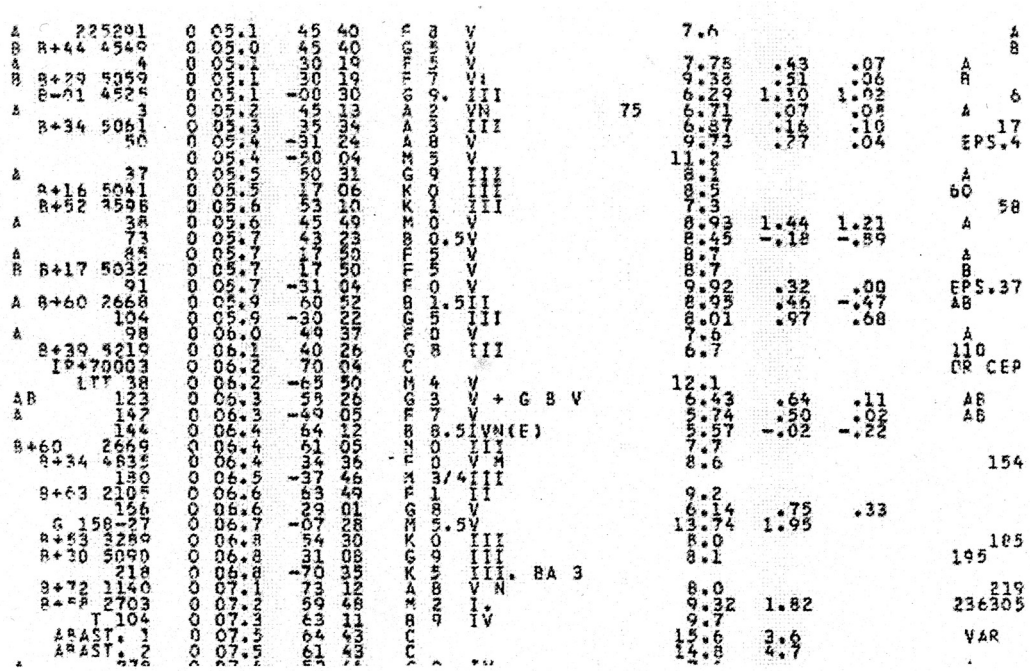
Since the key persons of this catalog project have been deceased, a systematic compilation

of these catalogs and supplements are looked forward by many astronomers.

III. *Backgrounds of developing OCR*

A sample page of the printed catalog is shown in Fig.1.

FIG. 1. SAMPLE PAGE OF THE TARGET CATALOG



The catalog contains a table of 227 pages, with 19189 rows and 80 columns. The meaning of each field is explained in Tab.1 which is cited from the so-called "ReadMe" file of the catalog archive system in CDS.

Judging from the face, the table seems to be printed by a kind of line-printer which was widely used in those years. The characteristics are:

1. Types are restricted to only upper case characters and numbers, with several symbols (".", "−", "+", "*", "/", "(", ")", ":").
2. Different fonts or different size of types are not used.
3. Column arrangement is very regular. Left-right fluctuations of types seem to be less than about 1/4 of type width. Row arrangement is in the similar situation.
4. Defects of types are often seen. Those of lower part of types cause difficulties in distinguishing "B", "P", and "R".
5. The sheet is enough clean.

TABLE 1. BYTE-BY-BYTE DESCRIPTION OF CATALOG FIELDS

Bytes	Format	Unit	Label	Explanation
1- 11	A11	---	Id	Identification for star
14- 15	I2	h	RAh	Right Ascension (2000) hour
17- 20	F4.1	min	RAm	?Right Ascension (2000) minute
23	A1	---	DE-	?Declination sign
24- 25	I2	deg	DEd	?Declination (2000) degree
27- 28	I2	arcmin	DEm	?Declination (2000) minute
29	A1	---	u_DEm	[:] DEm uncertain
31- 51	A21	---	SpType	MK type
52	A1	---	n_SpType	[*] ' * ' for a standard
53- 58	F6.3	mag	Vmag	? Johnson V magnitude
59- 63	F5.2	mag	B-V	? Johnson B-V color
65- 69	F5.2	mag	U-B	? Johnson U-B color
74- 80	A7	---	Name	Alternate identification, usually HD (1)

The OCR softwares available at present are difficult to use for this purpose. We tried "Table OCR for Excel v4.0" produced by FUJITSU, which always responded a message "Identification unsuccessful". Other general purpose OCR softwares are difficult especially in identifying column positions.

The published catalog is bound by a plastic binder, and is easily unbound. A scanner with excellent sheet-feeder is now available, and there is no problem in digitizing images of a lot of pages. We used "Scan Snap Color Image Scanner" produced by FUJITSU.

The scanner can adjust the column vertical, and provide an image file with JPEG form. We convert the JPEG to FITS form, by using a software "ImageMagick".

Were it not for these new tools, we would not have begun this project.

IV. *Development of OCR in Detail*

The procedure of character identification can be divided into following steps:

1. Determine top, bottom, left, and right edge of the printed area.
   1-1. If necessary, the rotation angle of the area should be determined.
   1-2. It is possible that the angle between column and row is not a right angle.
2. Determine scale factors, i.e., determine the numbers of dots per character for ordinate and abscissa.
   2-1. If necessary, dependence of scale factors with coordinates (i.e., distortion) should be determined.
3. Determine coordinates of origin point for each box area of a character.
4. Cutout a minimum rectangular area of the character type in the box area.
   4-1. It is possible that the printed position of the type shifts over the boundary of the box area.
5. Compare the cutout pattern with a list of type patterns, and identify the character.
   5-1. It is necessary to prepare several patterns of type for a character.

Details and actual results of these steps are explained below.

1. Determine four edges

We have scanned the page with the default resolution of the scanner, i.e. 144 dots per inch. In this case, a page is scanned as 1240x1752 dots image file. A dot has 1 byte gray scale, and we adopted 0 scale for perfect white, 255 for perfect black.

Calculating each sum of dots' data along column and row direction, we can see sharp edges of four sides of the printed area, because the rotation angle of the printed area from the rectangular coordinate are negligibly small for all pages. Thus determining coordinates of the four edges is not so problematic and the results are regular.

A slight inclination of row is found in some pages; the right end of the top edge is lower than the left end, by about one half of a character hight, though the deflection of column from vertical line is negligible. We determine this correction by comparing every line-space coordinate, between those of left one-third block of columns and those of right one-third one. Because the right block contains many blank lines, we cannot always find its top edge. We determine the line level by measuring coordinates of many line-spaces, developing a little complicated procedure.

Note that finding the left edge of the printed area is also difficult because the types in the first column are sparse. We use the right edge of column 11 instead the leftmost column.

2. Determine scale factors

The scale factors (i.e., numbers of dots per a character for ordinate and abscissa) are simply determined by dividing the width and hight of the whole printed area by respective character numbers (i.e., 80x83). The results are about 13.2 and 16.6 dots per a character, width and hight respectively. The scale factors are almost same for all pages, except several pages in which about 2% different factors are found.

If there is any distortion mechanism in printing processes, the scale factors differ depending the coordinates. We can estimate the differences by measuring every line-space coordinate as above, and column-space one.

In y-direction (i.e., ordinate direction) we find only a small difference, about ± 3 dots and not systematic. We apply no systematic corrections.

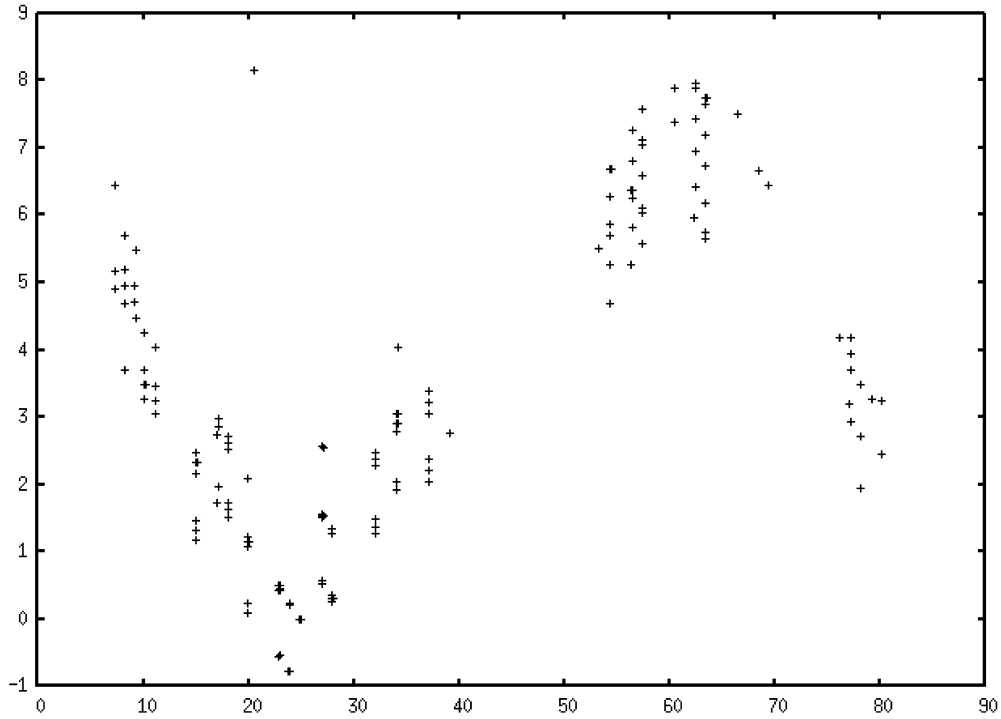
On the other hand, significant inhomogeneity is found in x-direction (i.e., abscissa direction) scale factor. It is shown in Fig.2.
The x-scale factor between columns 24 and 62 is different from the other. This is about 13.4 while that of the other part is about 12.9 dots per character. We apply this correction.

3. Determine the coordinates of origin point of a box area for a character

Given a character position with its column and row number, we calculate the coordinates of its origin point, i.e. the pixel dot coordinates of the top and leftmost point of a box containing the character. The box size is 14x17 dots. Applying those corrections described above, we can always catch a part of the character type at the central point of the box.

However the box area also contains parts of types of nearby characters. Thus we need the next procedure of cutout the type of the target character only.

FIG. 2. INHOMOGENEITY IN X-DIRECTION SCALE FACTOR



The abscissa denotes column number and ordinate denotes the displacement of column-spaces from the average coordinate. The difference of inclination means the difference of the scale factor.

4. Cutout a minimum rectangular area of the character type in the box area

We have tested several algorithms for this purpose. We finally adopt following algorithm.

First, we calculate correction of line-space coordinate, using the line data which contain the target character and several neighboring characters. This method causes very few exceptions and the upper and lower edge of the type are successfully determined.

Then we determine the left and write end of the type, searching them from the center to left or to right respectively.

This process seems to be the most important and complicated one in the OCR software, because present general OCR softwares often miss separating characters.

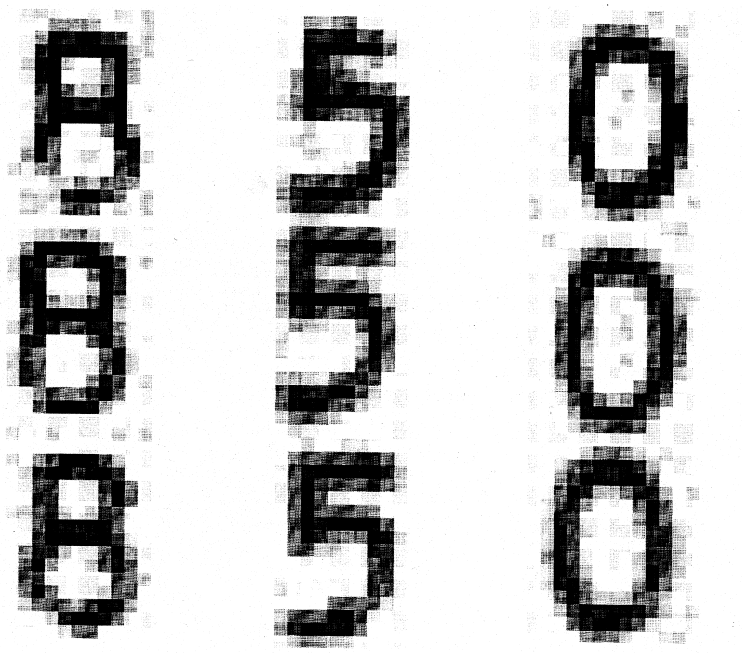
5. Compare the cutout pattern with a list of type patterns, and identify the character

We also have tested various algorithms in this case. The most successful one is that of computing the least square difference between the cutout pattern and several type patterns of each character.

Before comparing, we determine a threshold of distinguishing the type from the white background. We have tested several values and adopted about 1/5 of full-scale.

In order to construct a list of standard type patterns, we have inspected many printed types, and selected three different patterns for each character including those with defects. Example are shown in Fig.3.

FIG. 3. EXAMPLES OF THREE PATTERNS OF TYPES



Shifting the cutout pattern one dot by one dot over the standard type patterns, and computing the square sum of every difference, we determine the identified character, as that which gives the minimum square sum value.

The overall results are described and discussed in next section.

V. *Results and Discussions*

The rate of misidentification by our method is about 1.5% for identifying all characters in the catalog, though it is yet a rough estimate in preliminary test. If we restrict the identification within numeric types including "-" and ".", the rate is 0.15%, i.e. almost only two numeric characters in a page. An example of misidentification is shown in Fig.4.

Besides these misidentifications, there are some erroneous readings, identifying blots on the sheet as a character. These can be easily corrected by eye-inspecting, or by the consistency check of the data by a software. Also the blots can be erased from the digitized page images, by image tools such as "gimp".

We can attain a higher performance of identifications, dividing the page into several

FIG. 4. AN EXAMPLE OF MISIDENTIFICATION

Upper-right "8" is identified as "3", and lower-right "8" as "5".

blocks, depending that the characters are only numeric or not. Some non-numeric fields might be less important than numeric ones. For example, the top field denotes the name of the star, which can be reproduced from the star position and magnitude values referring other catalogs. It can be prolonged to check this field, until the time of final re-compilation of whole series of these catalogs. See next.

Using a computer of Intel Architecture with 1.4GHz CPU, the time for processing 1 page is less than 2 seconds. From the point of view of computer processing time, there is still room to increase the number of standard type patterns for one character (i.e., 5 or over instead of present 3) in order to improve the identification performance.

However, the main bottle-neck of the total performance is the human-eye inspection for checking misidentifications. It takes at least 10 minutes per one page, that is, 40 hours in total.

VI. *Conclusions and Future Works*

Judging from the performance of our OCR system, it is worth trying to compile the Buscombe's 6th catalog into machine-readable form by our system. We have begun this project from September 2004, and students are now engaging to check the results, at the time of writing this work.

As described in section 2, all the catalogs and supplements are waited to be re-compiled into one comprehensive catalog. Because there still remains a lot of errors or disorders in these catalogs and supplements, we must correct them at the time of re-compilation. We may need another system of software for the consistency check of the data, e.g. the cross check of star-positions and names, etc.

Though this catalog of stellar spectral classifications is rather old, we will be able to use this archive to find some changes in the "starry heavens", by comparing this with new catalogs

or observations.

HITOTSUBASHI UNIVERSITY

REFERENCES

Buscombe, W. 1977-2001, "MK spectral classifications. Third — Fifteenth general catalogue" (Northwestern Univ., Evanston, Illinois)
Gulyaev, A.P. & Nesterov, V.V. 1992, "On the four million stars catalogue" (Moscow University Press, Moscow) (in Russian)
Jaschek, C., Conde, H. & de Sierra, A.C. 1964, "Catalogue of Stellar Spectra Classified in the Morgan-Keenan System" (Astronomical Observatory of National University of La Plata)
Jaschek, C. 1978, Bull. Inform. CDS, 15, 121
Jaschek, C. 1989, Data in astronomy (Cambridge University Press, Cambridge)
Kennedy, P.M. & Buscombe, W. 1974, "MK Spectral Classification (1963-1973)" (Northwestern Univ., Evanston, Illinois)
Kennedy, P.M. 1983, "MK Classification Catalogue Extension" (Mt Stromlo & Siding Spring Observatories, Australia)
Nakajima, K. 1993, The Hitotsubashi Review, 110, 343 (in Japanese)
Nakajima, K. 1994, Hitotsubashi University Research Series Science, 29, 3 (in Japanese)
Urban, S.E., et al. 1997, "The AC2000: the Astrographic Catalogue on the Hipparcos System" (US Naval Observatory)