ASTRONOMY & ASTROPHYSICS
SUPPLEMENT SERIES

Astron. Astrophys. Suppl. Ser. **142**, 339–345 (2000)

Automatic extraction and classification of low-dispersion objective prism stellar spectra

E. Bratsolis[1,2], **I. Bellas-Velidis**[1], **A. Dapergolas**[1], **E. Kontizas**[1], **and M. Kontizas**[2]

[1] Astronomical Institute, National Observatory of Athens, P.O. Box 20048, GR-11810 Athens, Greece
[2] Section of Astrophysics, Astronomy and Mechanics, Department of Physics, University of Athens, GR-15784 Athens, Greece

Received July 22; accepted November 16, 1999

Abstract. The observing material used for this task is prism spectral plates taken with Schmidt-class telescopes. Such a plate generally contains thousands of spectra, and there are prism-plate libraries and digitized databases in several astronomical centers that can be exploited for this analysis. After a successive detection from the prism plate image, the spectra are automaticaly extracted in one-dimensional streams containing all the basic information. These spectra require automated classification methods to be analyzed in an objective form. In this article we compare two classification methods directly applied to stellar spectra: a linear correlation and a minimum distance method.

Key words: methods: data analysis — stars: fundamental parameters

1. Introduction

Stellar spectral classification is not only a tool for labeling individual stars but is also useful in studies like stellar population synthesis. Extracting the physical quantities from the digitized spectral plates involves three main stages: detection of the spectra, extraction of their images and classification of the spectra. The detection problem and their resolution for digitized objective prism Schmidt plates was presented by Bratsolis et al. (1998). The purpose of this paper is to present a fully automated method for the extraction and classification of spectra.

High-quality film copies of IIIa-J (broad blue-green band) plates taken with the 1.2 m UK Schmidt Telescope in Australia have been used. The spectral plates are with dispersion of 2440 Å/mm at H_γ and spectral range from 3200 to 5400 Å. The photographic material has been digitized at the Royal Observatory of Edinburgh using the Super-COSMOS measuring machine.

A classification problem can be formalized as a pair $(\mathcal{O}, \mathcal{C})$ where $\mathcal{O}$ denotes a set of objects and $\mathcal{C}$ a collection of disjoint subsets $\mathcal{C}_1, ..., \mathcal{C}_l$ that partitions $\mathcal{O}$. The problem is to determine the subset $\mathcal{C}_j \subset \mathcal{C}$ to which a given object $o \in \mathcal{O}$ belongs. In practice, the set of objects is usually very large, and providing an explicit description of each subset is impractical. The subsets are therefore often implicitly described by specifying a number of typical examples for each subset. Modern computational techniques have been developed to classify large databases of spectra, uniformly and in a considerably short time.

Automated classifiers of stellar spectra have been used in the past. Cross-correlation and minimum distance methods have been originally used by Kurtz (1982, 1984). A good review of linear multivariate statistical methods can be found in Murtagh & Heck (1984). Stellar classification with artificial neural networks (ANN) as a non-linear technique has been used by many other researchers during the last decade (von Hippel et al. 1994; Gulati et al. 1994; Vieira & Ponz 1995; Singh et al. 1998; Bailer-Jones et al. 1998). These methods were utilised for different databases and different spectral dispersion images.

The final stage of our work contains the stellar population synthesis of Magellanic cloud regions with a fully automated method. The detection procedure (Bratsolis et al. 1998) gives the stellar coordinates on the prism plate. This means that after the classification, we will have a complete mapping of different aged stellar groups of the studied region and this will be the subject of a future study. The low-dispersion objective prism plate was choosen to limit the overlaps between adjacent spectra. We discribe here the extraction method and we compare two simple and effective linear techniques of classification using a sample of 426 low-dispersion extracted stellar spectra from our digitized objective prism plate.

Send offprint requests to: E. Bratsolis
e-mail: ebrats@atlas.uoa.gr

2. Classification with the low-dispersion prism P1

An objective prism acts as a disperser through the effect of differential refraction. That is to say, the refractive index of the material of which it is composed varies with wavelength, and so rays of differing wavelengths are deviated to different extents on passage through the prism. The full-aperture objective prism P1 has an apex angle of 44 arcmin (2400 Å/mm at 4300 Å). The prism can be mounted on the telescope, and the complete assembly can be rotated to give the dispersion direction in any required position angle. The usual "default option" is to have the dispersion North-South. The dispersions and effective resolutions for the prism P1 are shown in Table 1.

Table 1. Dispersions and effective resolutions for the prism P1

Feature	λ(Å)	Dispersion (Å/mm)	Eff. Res. (Å)
H_α	6563	8088	176
H_β	4861	3470	76
H_γ	4340	2440	53
H_ϵ	3970	1852	41

The error in spectral classification of stars is likely to be approximately one spectral class. *O* and *B* spectra are not easily separated by a human classifier, so the P1 dispersion objective prism stellar classification contains the six following classes (Nandy et al. 1977; Krug et al. 1980; Cooke et al. 1981; Savage et al. 1985).

- *OB spectra*
  Uniform continuum intensity distribution over the length of the spectrum. The late *B* spectra show Balmer discontinuity at the UV region of continuum.
- *A spectra*
  Uniform continuum intensity distribution up to a strong Balmer jump at 3700 Å. Broad H_γ absorption line at 4340 Å. The absorption line H_β at 4861 Å is evident as a slight narrowing of the spectrum width.
- *F spectra*
  Intensity decreases slowly with decreasing wavelength. The only evident absorption is a blend of H_γ and *G*-band at 4340 and 4300 Å respectively. Calcium H and K lines at 3970 and 3934 Å respectively are not seen. This distinguishes *F* stars from *G* stars.
- *G spectra*
  Intensity decreases with decreasing wavelength. There is absorption at a strong blend of H_γ and *G*-band at 4340 and 4300 Å respectively. Absorption appearance at Calcium H and K lines at 3970 and 3934 Å respectively. Usually a *G* spectrum is shorter than an *F* spectrum.
- *K spectra*
  Rapid decrease in intensity with decreasing wavelength. The majority of spectral energy is between wavelengths 5400 to 4300 Å. There is a broad absorption line of CaI at 4227 Å. Absorption appearance at Calcium H and K lines at 3970 and 3934 Å respectively that is presented as a broad absorption at 3950 Å.
- *M spectra*
  Very rapid decrease in intensity with decreasing wavelength. A comparison between the *K* and *M* examples shows a more rapid decrease in the *M* star. Absorption line of CaI at 4227 Å. Absorption appearance of TiO bands at 5000 and 4800 Å.

3. Image reduction

Our image contains, in pixel size, a region of 3150(SN) × 3200(EW) of SMC. The scanning pixel size of a SuperCOSMOS measuring machine is 10 μm and the plate scale is 67.11 arcsec/mm. Our image with center $RA_{2000} = 1^h16^m$ and $DEC_{2000} = -73°20'$ contains a region of 35.2 arcmin(SN) × 35.8 arcmin(EW) of SMC (Fig. 1).

The saturation limit of the plate has a density value of 2.2. This means that for the pixels having negative decimal logarithm of density value greater than 2.2, we have nonlinear distortion of density-intensity plate diagram. The magnitude limit for classifying the spectra of our plate is $m_B = 18.5$.

4. The spectrum detection procedure DETSP

The spectrum detection procedure DETSP (Bratsolis et al. 1998) takes as input an image frame from the digitized spectral plate with particular parameters for the spectral images (dispersion, length, width). The subframe overlapping can be applied to both axes (defaulted), only down the wavelength axis, or skipped altogether. The processing is carried out in four sequential stages:

a. Image frame preprocessing
The whole image frame is filtered by a sequence of median and smoothing filters. A grid of subframes is fixed on the filtered image, according to the overlapping mode.

b. Subframe signal processing
Each one of the fixed subframes is processed by applying the detection algorithm based on a signal processing method. The detected spectral positions are saved in a table format.

c. Detection table processing
There are possible double detections of spectra near the edges of neighbouring subframes. For this reason, the table of detected spectra is now processed to remove the doubling. It is sorted as well.

d. Detections fine adjustment
The signal processing approach is used again. As many subframes are fixed as the number of detected spectra. The subframes are narrower and each one includes a particular detected spectrum image. This leads to fine adjustment of the position. The adjusted position table is finally sorted.

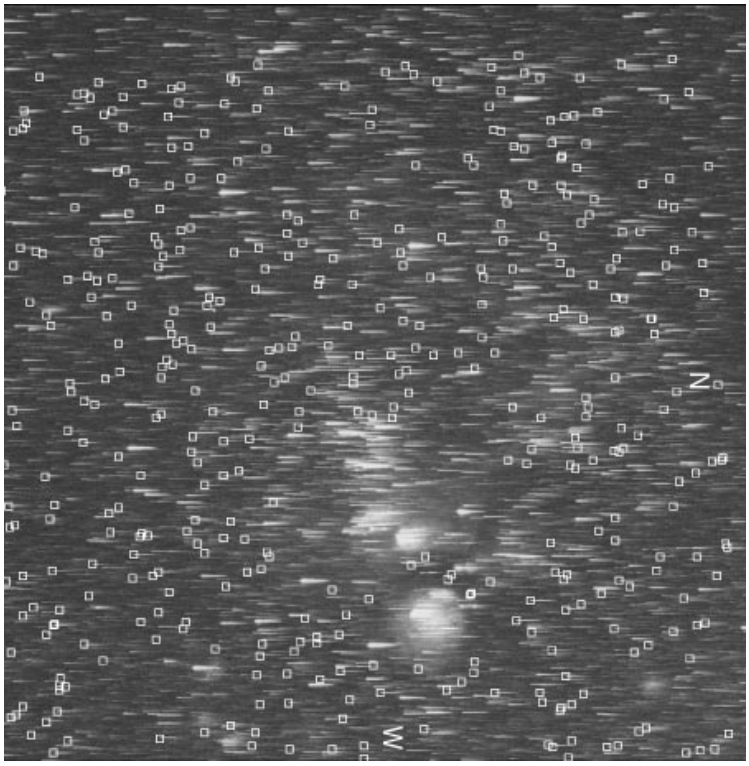

Fig. 1. Super-COSMOS negative image of a region 35.2 arcmin(SN) × 35.8 arcmin(EW) of SMC

5. Spectral extraction

One of the advantages of the Super-COSMOS machine is that it scans the plates with a direction parallel to the longtitudinal axis of the spectra. Thus, our spectra are parallel to a coordinate axis. The success of the DETSP procedure is that it detects all the spectra at the same common-wavelength zero-point at 5400 Å. This zero-point (0.000 mm) corresponds to our pixel scale $(1 - 128)$ at 10 pixels.

After the spectral detection, a new procedure starts, responsible for the extraction of spectra (EXTSP). The spectral length contains 128 pixels. These are: the zero-point plus 118 pixels on the right of zero-point plus 9 pixels in the left of zero-point. For a better signal-to-noise ratio, the actual extraction of the spectrum is performed by means of rectangular weighted "slit" sliding on data (Balestra et al. 1990). Its width and shape are either fixed or determined by the average fit on the transversal sections of the spectrum.

The new zero-point defined by DETSP at 5400 Å had to be added to the dispersion curve of the objective prism P1 (Nandy et al. 1977). The parallel displacement in mm of zero-point gives new distance measurements for various features. The results are shown in Fig. 2 and Table 2.

The extracted spectra are stored in a two-dimensional file $n \times 128$, where $n = 426$ is the number of detected spectra. Every row of this file is an independent normalized spectrum with length 128 pixels (Fig. 3).

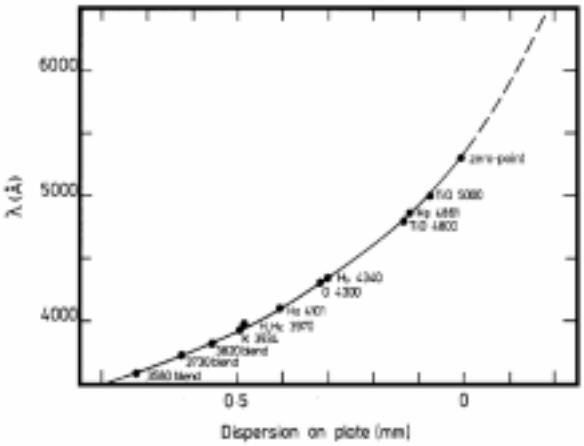

Fig. 2. Dispersion curve for objective prism P1

6. Classification by use of linear correlation

The most widely used method to measure the dependence on two variables, is the linear correlation coefficient r.

Table 2. Details for the features on objective prism P1

Feature	λ(Å)	Distance (mm)	Pixel Num.
Zero-Point	5400	0.000 ± 0.005	10 ± 1
TiO	5000	0.100 ± 0.005	20 ± 1
H_β	4861	0.150 ± 0.005	25 ± 1
TiO	4800	0.160 ± 0.005	26 ± 1
H_γ+G	4340, 4300	0.320 ± 0.005	42 ± 1
CaI	4227	0.340 ± 0.005	44 ± 1
H_δ	4101	0.430 ± 0.005	53 ± 1
$H+H_\epsilon$	3970	0.500 ± 0.005	60 ± 1
CaII+K	3936, 3934	0.520 ± 0.005	62 ± 1
MgI+FeI blend	3820	0.570 ± 0.005	67 ± 1
FeI+H blend	3730	0.640 ± 0.005	74 ± 1
FeI blend	3580	0.740 ± 0.005	84 ± 1

Fig. 3. A sample of 426×128 spectra

The spectra are normalized and they are considered as vectors in R^N with $N = 128$.

Let $D_{ij} = D_i(\lambda_j); j = 1, ..., 128$ be the normalized density value for the i^{th} stellar spectrum and $S_{kj} = S_k(\lambda_j)$; $j = 1, ..., 128$ be the normalized density value of the k^{th} class standard stellar spectrum with $k = 1, ..., 6$. For $k = 1, ..., 6$ the standard stellar spectra are $OB, ..., M$ of Figs. 4-9. The correlation coefficient for the i^{th} stellar spectrum for the k^{th} class is

$$r_{ik} = \frac{\sum_{j=1}^{128}(D_{ij} - \bar{D}_i)(S_{kj} - \bar{S}_k)}{\sqrt{\sum_{j=1}^{128}(D_{ij} - \bar{D}_i)^2}\sqrt{\sum_{j=1}^{128}(S_{kj} - \bar{S}_k)^2}}, \quad (1)$$

with $\bar{D}_i$ being the mean value (over the j variable) of the i^{th} spectrum, $i = 1, ..., 426$ and $\bar{S}_k$ the mean value of the k^{th} class standard spectrum.

The correlation coefficient r_{ik} for the i^{th} spectrum for the class k was calculated with displacement ± 3 pixels to predict a possible displacement from the detection algorithm caused by the local background. For these seven correlation coefficients for every class k, the maximum value was choosen. The final classification was given by the maximum value of the coefficient r_i of all the r_{ik} coefficients as

$$r_i = \arg\,(\max\ \ r_{ik}), \quad k = 1, ... 6. \quad (2)$$

7. Classification by use of minimum distance

The normalized spectra were standardized and a metric was introduced in vector space (Murtagh & Heck 1984; Vieira & Ponz 1995). After standardization, we obtain

$$D_{ij}^{\text{new}} = \frac{(D_{ij}^{\text{old}} - \bar{D}_i)}{\sigma_i} \quad (3)$$

with $\bar{D}_i$ being the mean value (over j variable) and σ_i the standard deviation of the i^{th} spectrum; $i = 1, ..., 426$. The spectra have zero mean and unit standard deviation. The metric is the standard unweighted Euclidean distance between two real-valued vectors d_{ik} given by

$$d_{ij}^2 = \sum_{j=1}^{128}(D_{ij} - S_{kj})^2. \quad (4)$$

The minimum distance d_{ik} for the i^{th} spectrum for the class k was calculated with displacement ± 3 pixels to predict a possible displacement from the detection algorithm caused by the local background. The final result was given by

$$d_i = \arg\,(\min\ \ d_{ik}), \quad k = 1, ... 6. \quad (5)$$

8. Results and discussion

The field studied here contains the associations NGC 456, NGC 460 a,b and NGC 465 of SMC. These associations occupy only 10% of the studied region. The majority of the stellar spectrum owing to the associations, superimposed to HII regions, has been saturated and excluded from our

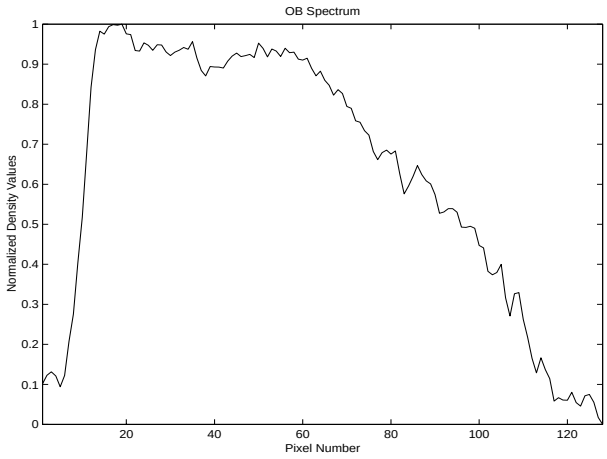

Fig. 4. A characteristic 1×128 OB spectrum of our sample

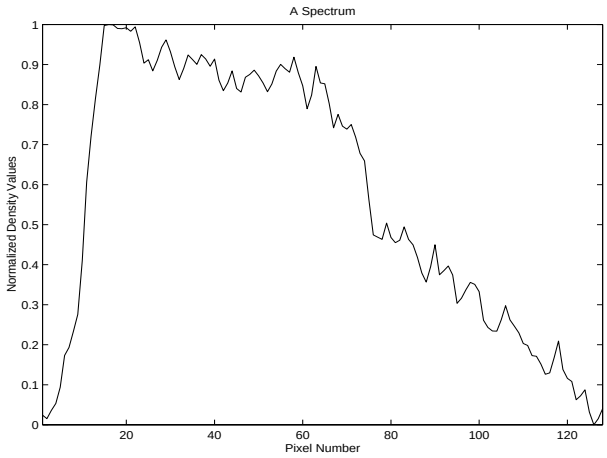

Fig. 5. A characteristic 1×128 A spectrum of our sample

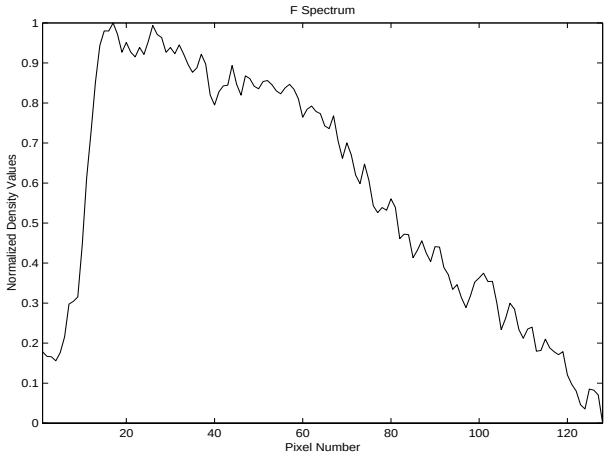

Fig. 6. A characteristic 1×128 F spectrum of our sample

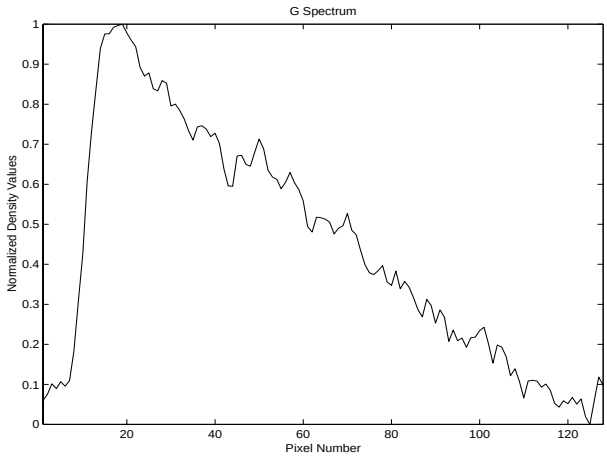

Fig. 7. A characteristic 1×128 G spectrum of our sample

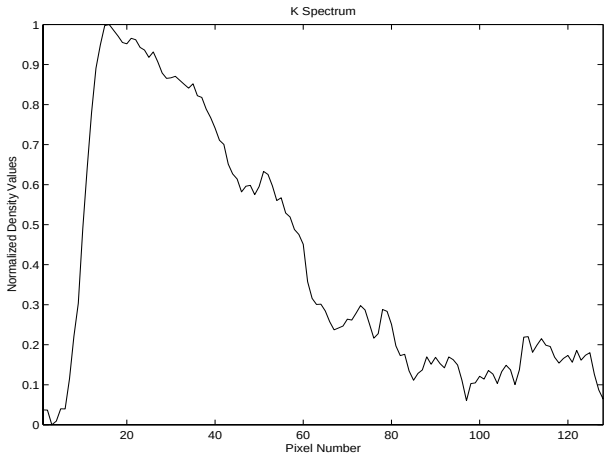

Fig. 8. A characteristic 1×128 K spectrum of our sample

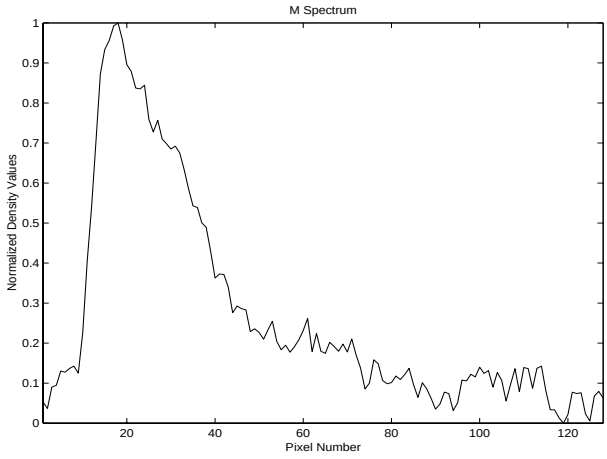

Fig. 9. A characteristic 1×128 M spectrum of our sample

sample. The accuracy of this classification is ±1 spectral type (Dapergolas et al. 1991).

The method has been tested on spectra for which a visual classification was available. The down limit for the faint spectra was the same as in visual classification. For the bright spectra the limit has been described in Sect. 3. The method has been developed on photographic objective prism plates but it can be equally be well applied to CCD objective prism images.

The results from human expert (HE), linear correlation (LC) and minimum distance (MD) method are shown in Table 3. Figures 10-12 show the histograms for each method.

Table 3. Details for the different classification methods

Spectral Type	OB	A	F	G	K	M
Class	1	2	3	4	5	6
HE	150	63	19	47	63	84
LC	148	48	32	61	62	75
MD	145	41	38	64	65	73

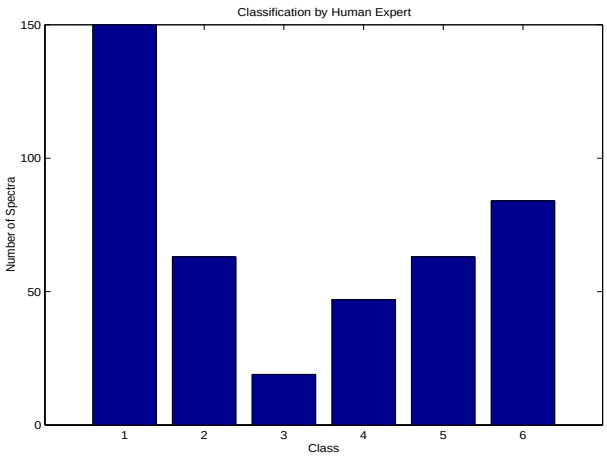

Fig. 10. Histogram for the human classifier

To quantify the degree of agreement between different classification methods we have calculated the mean error $me_{\rm hehe}$ between two human experts a and b, $me_{\rm helc}$ between human expert and linear correlation classification and $me_{\rm hemd}$ between human expert and minimum distance classification and the corresponding dispersions $\sigma_{\rm hehe}$, $\sigma_{\rm helc}$ and $\sigma_{\rm hemd}$ using the following equations

$$me_{\rm hehe} = \frac{1}{426}\sum_{i=1}^{426} \mid C^i_{\rm he(a)} - C^i_{\rm he(b)} \mid = 0.23$$

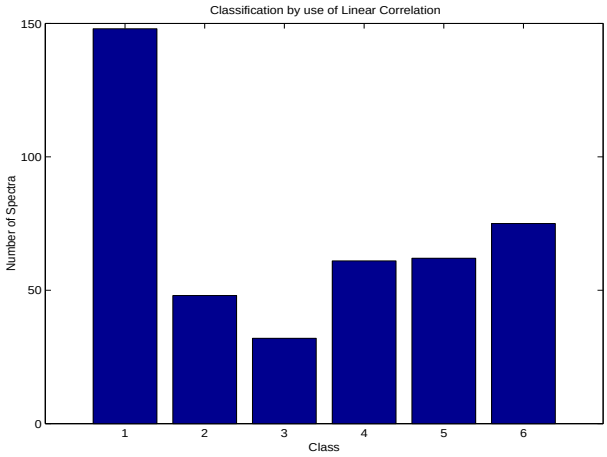

Fig. 11. Histogram for the linear correlation method

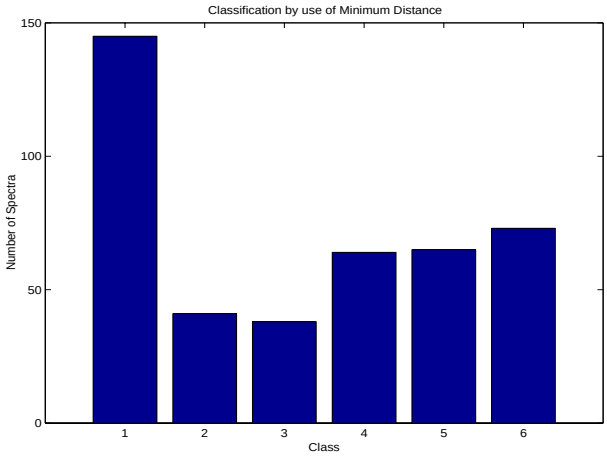

Fig. 12. Histogram for the minimum distance method

$$me_{\rm helc} = \frac{1}{426}\sum_{i=1}^{426} \mid C^i_{\rm he} - C^i_{\rm lc} \mid = 0.27$$

$$me_{\rm hemd} = \frac{1}{426}\sum_{i=1}^{426} \mid C^i_{\rm he} - C^i_{\rm md} \mid = 0.29$$

$$\sigma_{\rm hehe} = \sqrt{\frac{1}{426}\sum_{i=1}^{426}(C^i_{\rm he(a)} - C^i_{\rm he(b)})^2} = 0.47$$

$$\sigma_{\rm helc} = \sqrt{\frac{1}{426}\sum_{i=1}^{426}(C^i_{\rm he} - C^i_{\rm lc})^2} = 0.58$$

and

$$\sigma_{\rm hemd} = \sqrt{\frac{1}{426}\sum_{i=1}^{426}(C^i_{\rm he} - C^i_{\rm md})^2} = 0.60.$$

Table 4. Statistical properties for the different classification methods

Test	mean error	dispersion
hehe	0.23	0.47
helc	0.27	0.58
hemd	0.29	0.60

For comparison reasons, we display these results in Table 4. The two automated methods of classification seem to be close to the human expert classification for the low-dispersion prism (P1) stellar spectra with linear correlation giving better results. An extended study for classification with artificial neural networks is under preparation.

Acknowledgements. This research has been supported by a grant from the General Secretariat of Research and Technology of Greece, PENED program. The authors are grateful to the UK Schmidt Telescope Plate Library (ROE) for the loan of the observational material.

References

Bailer-Jones C.A.L., Irwin M., von Hippel., 1998, MNRAS 298, 361

Balestra A., Micol A., Pasian F., Santin P., Sedmak G., Smareglia R., 1990, in Proceedings of 2nd ESO/ST-ECF Data Analysis Workshop, Baade D., Grosbol (eds.), ESO Conf. and Workshop Proc., p. 36

Bratsolis E., Bellas-Velidis I., Kontizas E., Pasian F., Dapergolas A., Smareglia R., 1998, A&AS 133, 293

Cooke J.A., Emerson D., Kelly B.D., MacGillivray H.T., Dodd R.J., 1981, MNRAS 196, 397

Dapergolas A., Kontizas E., Kontizas M., Pasian F., Pucillo M., Santin P., 1991, A&AS 87, 97

Gulati R.K., Gupta R., Gothoskar P., Khobragate S., 1994, ApJ 426, 340

Krug P.A., Morton D.C., Tritton K.P., 1980, MNRAS 190, 237

Kurtz M.J., 1982, Automatic Spectral Classification, Ph.D. Thesis, Dartmouth College, New Hampsire

Kurtz M.J., 1984, in The MK Process and Stellar Classification, Garrison B.F. (eds.). David Dunlop Observatory, Toronto, p. 136

Murtagh F., Heck A., 1984, Multivariate Data Analysis. Reidel, Dordrecht

Nandy K., Reddish V.C., Tritton K.P., Cooke J.A., Emerson D., 1977, MNRAS 178, 63P

Savage A., Beard S.M., Palmer J.B., 1985, The UKST Objective Prisms I. Royal Observatory of Edinburgh

Singh H.P., Gulati R.K., Gupta R., 1998, MNRAS 295, 312

Vieira E.F., Ponz J.D., 1995, A&AS 111, 393

von Hippel T., Storrie-Lombardi L.J., Storrie-Lombardi M.C., Irwin M.J, 1994, MNRAS 269, 97