AUTOMATED SPECTRAL CLASSIFICATION AND THE GAIA PROJECT

Jerry LaSala[1], **Michael J. Kurtz**[2]

[1]Physics Department, University of Southern Maine, 96 Falmouth Street, Portland, Maine 04103, USA
[2]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

ABSTRACT

Two-dimensional spectral types for each of the stars observed in the GAIA mission would provide valuable additional information for galactic structure and stellar evolution studies, as well as aiding in the identification of unusual objects and populations. Classification of the enormous number of spectra needed for such a project makes automated techniques an absolute necessity. We present a brief survey of approaches to automatic classification, then discuss our Metric-Distance method, which in developmental tests produces spectral types with mean errors comparable to those of human classifiers working at similar resolution. We discuss data and equipment requirements for an automated classification survey. Finally, we propose a program of auxilliary observations to yield spectral types and radial velocities for the GAIA stars.

Key words: GAIA, spectral classification

1. INTRODUCTION

The spectral class of a star is a fundamental classical datum in stellar astronomy. The two-dimensional MK spectral type can provide provide information about the temperature and surface gravity of a star, through interpretation in light of theoretical models of stellar atmospheres. But spectral types are not classifications by temperature and gravity; they are descriptions of the visual appearance of the spectra. This is one of the most important characteristics of the MK spectral types and the MK process in general. The spectral types are model-independent; theoretical advances may change the temperature and surface gravity associated with spectral type A0V, but a star of type A0V will still be an A0V.

Moreover, from the very earliest era of spectral classification, the types have been based on the overall appearance of the spectra, and not on continuum shape, particular line strengths, or particular ratios of lines. The centrality of the model-independence and emphasis of total appearance has been emphasized by authors from Payne (1925) through Morgan (1984).

Standard MK classification involves comparing spectra to be classified with standard spectra of defined class. Ideal classification spectra have dispersions of about 67 Å/mm and cover a spectral range of approximately 3850–5000Å. Satisfactory classification is actually possible with spectra of about half that dispersion, e.g., the 112 Å/mm spectra used by Houk (Houk & Cowley 1975; Houk 1978, 1982). The minimum spectral resolution for reliable MK classification is about 1 Å.

One might ask 'Why do spectral classification at all? Can't we get the same information from multicolor photometry and even more information from high-resolution spectroscopy?' First of all, the MK process is a very powerful technique for using all the information in a stellar spectrum and integrating it with a unique perspective, which complements all of the other techniques. Other techniques give valuable, but different, information (see also Favata & Perryman 1995, Bastian 1995, for related complementary information being considered in the context of GAIA). For example, broad-band photometry looks at the deep photosphere, whereas the line spectrum is taken from several levels above the photosphere depending on the strength of the line and the part of the profile used. Similarly, classification has one big advantage when compared with quantitative, high-resolution spectroscopy. The problem of determination of the continuum and of equivalent widths is circumvented by the use of standard stars, so it is a very useful complementary check on equivalent-width methodology, as well as a useful source of new information not available in the quantitative techniques. Thus these various techniques are complementary rather than competitive.

2. AUTOMATIC CLASSIFICATION

It is clear that for the huge number of objects to be observed in the GAIA program, or any similar large-scale survey of fainter stars, that methods of automatic classification are the only feasible means of assigning spectral types. The desirability of a fully automated system of spectral classification has been stressed by many authors over the past two and a half decades (see, e.g. West 1973, 1976; Schmidt-Kaler 1979, 1982; Kurtz 1984; Houk 1976, 1984, 1994; Keenan 1987; and Garrison 1988). In addition to increasing the speed of the classification process, automated systems offer the prospect of greater durability and homogeneity: the ability to classify large numbers of stars on a self-consistent system. Large homogeneous samples are especially important for statistical studies, and are virtually unobtainable by traditional classification methods.

Table 1: Recent results in automated spectral classification

Source	Dispersion (Resolution)	Technique	Mean Error (Subtypes)
Kurtz (1982)	(14 Å)	Pattern Recognition	1.9
LaSala (1989)	112 Å/mm	Pattern Recognition	2.2
Kurtz & LaSala (1991)	112 Å/mm	Pattern Recognition	1.14
von Hipple et al. (1994)	112 Å/mm	Neural Net	1.7
Malyuto & Shvelidze (1994)	166 Å/mm	Quantitative	0.1
LaSala (1994)	67 Å/mm	Pattern Recognition	0.4
Weaver & Torres-Dodgen (1995)	(15 Å)	Neural Net	0.5

Other expected benefits of automated classification include the detection of variability, the possibility of classification in more than two dimensions, and the rapid detection of peculiar objects. Many consider this last the most important and promising contribution of automatic spectral classification and indeed of spectral classification in general: the classification process acts as a filter which allows one to select and study in great detail only the most interesting objects, or a few normal objects which are representative of a much larger group.

Various attempts to develop an automated system of spectral classification have been made over the past 25 years. Several developments within the past few years have finally converged to make a large-scale automatic classification program both necessary and feasible. The availability of high-speed, low noise plate scanners and, more recently, multi-object fiber-fed CCD spectrographs, has increased the rate of acquisition of spectral data far beyond the ability of human classifiers while providing data in ideal form for computerized analysis and classification. Simultaneously, increasingly fast and inexpensive computers and developments in the fields of image processing and pattern recognition have provided the mechanisms for a working automated classifier.

Reviews of previous work are given by West, Schmidt-Kaler, and Kurtz in the papers cited above, as well as by von Hipple et al.(1994). With these authors, we may divide techniques applied to automatic spectral classification into two categories which may be called quantitative methods and pattern recognition respectively. Quantitative methods involve the measurement of specified spectral quantities (equivalent widths of certain spectral lines, ratios of certain line strengths, etc.) and calibration of these measurements in terms of desired parameters such as spectral type and luminosity class.

Pattern recognition methods effect classification by determining a suitably defined similarity measure and assigning a given spectrum to the class whose standard the spectrum most closely resembles. Pattern recognition techniques involve direct comparison with standard spectra, while criterion evaluation methods do not; in this respect pattern recognition more closely resembles the process of visual classification. In addition, pattern recognition methods avoid the pitfalls of making absolute measurements of equivalent widths. Because of its closer correspondence with the visual classification technique, its use of the overall appearance of the spectrum, and its avoidance of model-dependent calibration, we believe strongly that pattern recognition is the approach which will ultimately yield the most powerful and most useful automated classification techniques.

There are two pattern-recognition approaches currently used by those developing automated spectral classification methods. These are Artificial Neural Networks (ANN) and Weighted Metric Distance algorithms. ANNs are a widely-used non-linear system based on a simple model of human neurons. The next speaker (Lahav 1995) will discuss ANNs in detail; in addition an excellent discussion of the application of ANNs to spectral classification may be found in Weaver and Torres-Dodgen (1995). The Metric Distance methods will be discussed extensively below. Both ANNs and Weighted Metric Distance methods base their classifications on weighted parameters determined from a learing set of classified spectra. We believe that the Metric Distance techniques are superior in that there are no 'hidden layers' involving complicated and often difficult to interpret relationships among the classification parameters. In addition, the Metric Distance methods allow direct comparison with standard spectra at the classification step, as in the traditional MK Process; ANNs do not.

Table 1 summarizes the most significant recent results in automatic classification. For comparison, mean errors for visual classification by trained classifiers range from about 1.0–1.3 subtypes (Houk & Cowley 1973) to as little as 0.44–0.63 subtypes (Houk 1978) for the top experts.

3. THE METRIC-DISTANCE ALGORITHM

Each digitized spectrum is regarded as an n-element vector, where n is the number of resolution elements or pixels. The standard spectra S represent fixed points in this space. The metric distance d_{xs} between a program spectrum X and a standard S is given by:

$$d_{xs}^2 = \frac{1}{n}\sum_{i=1}^{n} \alpha^2(i)[X_i - S_i]^2 \qquad (1)$$

where α is a weighting factor to be defined. If $\alpha(i)$ is set to 1, we have the ordinary Euclidean metric; Penprase (1993) uses a step function: $\alpha = 1$ at selected features, $\alpha = 0$ elsewhere.

Kurtz (1984) proposed defining:

$$\alpha^2(i) = \frac{\sum^2(i) - \sigma^2(i)}{\sigma^2(i)} \qquad (2)$$

where $\sum^2(i)$ is the variance of pixel i within the error box associated with some initial approximation, e.g. 'G',

and $\sigma^2(i)$ is the variance of the same pixel within the final classification box, e.g. 'G2'. This defintion gives the most weight to those features which have the smallest variance within the final classification box and larger variance among boxes. Thus it weights most strongly those features which discriminate the final classification from the general mix of spectra. We have adopted and continue to use this metric in our work.

4. RECTIFICATION

An important point to consider in any automated classification program is the question of eliminating the continuum and instrumental response. In a typical spectrum, the continuum and instrumental response represent first-order information, and the spectral lines second order information. MK classification is based on small differences in the strength and profiles of the lines, which is third-order information. For this reason, any residual errors in continuum removal will influence classification programs as much as or more than the third-order information on which classifications are to be based. In particular, since the continuum shape is known to be strongly correlated with the classification results (this is, after all, the basis of photometric classification), improper removal of this first-order information will bias the results toward the photometric classifications.

Figure 1. Top: B3V spectrum. Middle: B7II spectrum. Bottom: difference between top and middle spectra. All spectra have been 'reflattened' as described in text.

We have tested a number of continuum removal techniques and currently prefer a multi-step process we call 're-flattening'. We begin by applying the Fourier division technique (LaSala & Kurtz 1985) to produce a rectified spectrum. Then a Fourier-smoothed residual spectrum is calculated and subtracted from the rectified spectrum to produce the reflattened spectrum. In addition, anytime we calculate a difference between spectra, we reflatten the difference spectrum to remove any third-order continuum effects. As Fig. 1, from Kurtz & LaSala (1991), shows, the results are quite good.

5. AN OPPORTUNITY, AND A BURDEN

It is critically important to realize that a project of the scope being proposed, classification of approximately 50 million spectra of the stars observed by GAIA, will essentially define the spectral classification system of the future, as surely as Annie Cannon defined the Harvard types through classifying the 240 000 spectra of the Henry Draper Catalog. To insure that the new system thus defined is compatible with the existing body of MK classifications, extensive coordination with expert visual classifiers will be necessary, especially in the early stages of the program.

Some have suggested that we not worry about this, and simply define our new system without reference to the old, perhaps based on comparison with model stellar spectra. Such an approach is unwise, in that it discards both the existing body of classification research and methodology, and the model-independence which is an important virtue of the traditional methods.

Perhaps the spectra can be allowed to 'classify themselves' into natural groupings which may or may not correspond to the traditional MK types; this can be done using ANNs, Metric Distance algorithms, or other statistical methods. This is probably a good idea, but it must be done after, or in tandem with, a more traditional MK-anchored classification. Otherwise interpretation and analysis of the new classifications, and relating this new information to the existing body of knowledge, will be difficult or impossible.

6. INSTRUMENTATION

In the past, slitless spectrographs such as objective-prism telescopes were the principal source of spectra in such large volume that automatic classification methods would be of value. Today, the advent of multi-object fiber-optic spectrographs makes possible the high-speed acquisition of huge numbers of slit spectra.

For example, the 'Hectospec' spectrograph (Fabricant et al. 1994), currently being built for the soon-to-be-reconfigured Multiple Mirror Telescope, can observe 300 spectra simultaneously, and the fibers can be automatically repositioned for a new observation in 5 minutes. It seems quite feasible to build a similar device with 1000 fibers, capable of observing 1000 spectra simultaneously. A spectrograph of this design eliminates the source confusion (spectrum overlap) problems that arise with slitless spectra, since the position of the spectrum on the detector is fixed by the position of the output end of the optical fiber, not by the position of the star in the field. Coupled with a telescope of suitable aperature and wide field, such a spectrograph could obtain the 50 million spectra needed to complement the GAIA mission in a period of three years.

7. A MODEST PROPOSAL

We propose that auxilliary spectroscopy for the GAIA mission by performed not by a satellite-mounted telescope in tandem with GAIA, but by a dedicated earth-based facility.

A telescope of 4-m aperture, with a wide field of view, coupled with a 1000 fiber spectrograph of capability comparable to the Hectospec, would allow observation of stars down to 15 mag at a rate limited primarily by the repositioning time for the fibers. If used with a 1200 lines/mm grating, such a spectrograph would provide a

dispersion of about 0.3 Å/pixel over a spectral range of about 1000 Å. This is ideal for automated MK classification and would also allow radial velocity determination at a precision of 1 km/s. Indeed, such a high-stability spectrograph, with a signal-to-noise ratio appraoching 1000 for a 5-minute exposure of a 15 mag star, will allow radial velocity determinations of precision previously obtainable only with coude or echelle instruments.

The advantages of such an Earth-based program are many. All required instruments represent applications and extensions of existing technology. No special telemetry or new data reduction techniques are required. Source confuision is not an issue with a fiber-optic spectroscope. Construction costs are probably less than the development, construction, and increased launch costs associated with a space-based telescope. After the three years required to make the complete set of observations, repeat observations may be made to detect variability. And after the GAIA mission is over you still have a fully operational observatory with a first-class spectrograph.

8. THE PROMISE

As Hipparcos brought astrometry into the modern age, GAIA, coupled with the proposed auxilliary observations, can bring the study of stellar kinematics and populations into the new century.

ACKNOWLEDGEMENTS

JLS wishes to thank the Margaret Cullinan Wray Charitable Lead Annuity Trust, through the AAS Small Research Grants Program, and the NASA JOVE program for partial support for this work; also ESA for support for attendance at this workshop.

REFERENCES

Bastian, U., 1995, ESA SP-379, this volume

Fabricant, F., Hertz, E. H., Szentgyorgi, A. H. 1994, in Instrumentation in Astronomy VIII, ed. D. Crawford, Bellingham: SPIE, p. 251

Favata, F., Perryman, M.A.C., 1995, ESA SP-379, this volume

Garrison, R. F. 1988, PASP, 100, 1036

Houk, N., Cowley, A.P., 1975, University of Michigan Catalog of Two-Dimensional Spectral Types for HD Stars, Vol. 1 (Ann Arbor: University of Michigan)

Houk, N., 1978, 1982, University of Michigan Catalog of Two-Dimensional Spectral Types for HD Stars, Vols 2, 3 (Ann Arbor: University of Michigan).

Houk, N. 1976, remarks following West (1976)

Houk, N. 1984, remarks following Kurtz (1984)

Keenan, P.C. 1987, PASP, 99, 713

Kurtz, M.J. 1982, Thesis, Dartmouth College

Kurtz, M.J. 1984, in The MK Process and Stellar Classification, ed. R.F. Garrison (Toronto: David Dunlap Observatory) p. 131

Kurtz, M. J., LaSala, J. 1991, in Objective-Prism and Other Surveys, ed. A. G. D. Phillip and A. Upgren (Schenectady, NY: L. Davis Press) p. 133

Lahav, O., 1995, ESA SP–379, this volume

LaSala, J. 1989, Bull. AAS, 21, 759

LaSala, J. 1994, in The MK Process at 50 Years, ed. C.J. Corbally, R.O. Gray, and R. F. Garrison, ASP Conference Series, Vol. 60, p. 312

Malyuto, V., Shvelidze, T . 1994, in The MK Process at 50 Years, ed. C.J. Corbally, R.O. Gray, R.F. Garrison, ASP Conference Series, Vol. 60, p. 344

Morgan, W.W. 1984, in The MK Process and Stellar Classification, ed. R.F. Garrison (Toronto: David Dunlap Observatory) p. 18

Payne, C. 1925, Stellar Atmospheres (Cambridge: Harvard College Observatory)

Penprase, B. E. 1994, in The MK Process at 50 Years, ed. C.J. Corbally, R.O. Gray, R. F. Garrison, ASP Conference Series, Vol. 60, p. 325

Schmidt-Kaler, Th. 1979, in I.A.U. Colloquium 47: Spectral Classification of the Future, eds. M.J. McCarthy, S.J., A.G.D. Philip, G.V. Coyne, S.J.: Ricerche Astronomiche, Vol. 9, p. 285

Schmidt-Kaler, Th. 1982, Bulletin d'Information du Centre de Donnees Stellaires, Vol 23, p. 2

von Hipple, T., Storrie-Lombardi, L., Storrie-Lombardi, M. 1994, in The MK Process at 50 Years, ed. C. J. Corbally, R. O. Gray, R.F. Garrison, ASP Conference Series, Vol. 60, p. 289

Weaver, W.B., Torres-Dodgen, A.V. 1995, ApJ, 446, 300

West, R.M. 1973, in I.A.U. Symposium 50: Spectral Classification and Multicolor Photometry, eds. Ch. Ferenbach and B.E. Westerlund (Dordrecht: Reidel) p. 109

West, R.M. 1976, in Proceedings of the Third European Astronomical Meeting, ed. E.K. Karadzi (Tbilissi: Abastumani Astrophysical Observatory) p. 23