Chin. J. Astron. Astrophys. Vol. 1, No. 6, (2001) 563–572
(*http://www.chjaa.org* or *http://chjaa.bao.ac.cn*)

Chinese Journal of
Astronomy and
Astrophysics

Steps Towards a Fully Automated Classification and Redshift-measurement Pipeline for LAMOST Spectra. I. Continuum level and wavelength estimation for galaxies

A-Li Luo ⋆ and Yong-Heng Zhao

National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012

Received 2001 June 19; accepted 2001 July 12

Abstract The Large Sky-Area Multi-Object Spectroscopic Telescope (LAMOST) under construction by the National Astronomical Observatories will yield up to four thousand multi-fiber spectra of stars and galaxies per field. The present series of papers describes the automated data-reduction pipeline currently being designed in order to cope with the anticipated flood of spectrographic data. In this preliminary paper, we present an automated method for estimating the continuum level, the positions of strong lines and the 4000 Å break in galaxy spectra. In order to obtain detailed information on the continuum, we use a wavelet filter bank. After continuum fitting, our software searches for a 4000 Å break and distinguishes between emission-line galaxies (ELGs) and non-ELGs according to whether the break is small or large. It then searches for strong lines and measures the intensities of emission lines and the equivalent widths of absorption lines. For non-ELGs, the absorption lines are identified automatically yielding redshift measurements.

Key words: methods: data analysis — techniques: spectroscopic — galaxies: emission lines – galaxies: absorption lines

1 INTRODUCTION

There are already several wide-field multi-fiber spectrographic surveys in various stages of development, most notably the stellar population, galaxy and QSO surveys using the 2dF facility as described by Lewis et al. (1998) and the Sloan Digital Sky Survey (SDSS) as summarized by York et al. (2000), both of which are already underway, and those surveys planning to use the 6dF facility as described by Watson et al. (2000) and the LAMOST facility as outlined by Wang et al. (1996). These surveys are yielding or will yield very large numbers of fiber spectra. Automated methods for reducing the spectra (including the reduction of 2D images to 1D spectra) have therefore had to be developed.

⋆ E-mail: lal@lamost.bao.ac.cn

The SDSS spectroscopic data are reduced automatically by a pipeline which extracts, corrects and calibrates the spectra, determines the spectral types and measures the redshifts (York et al. 2000). It uses a wavelet method when necessary in order to deblend close lines. As the SDSS reduction software is based on IRAF routines with minor modifications, we study the IRAF spectrum-reduction software. Using IRAF, one needs to select interactively spectral regions where there are no strong lines or bands. The continuum is fitted in these selected regions. Then, one needs to identify some strong lines visually and mark them, which requires some experience on the part of the user. After using the positions of these lines to compute a redshift, IRAF then identifies other comparison lines and absorption lines automatically.

For the LAMOST project, how best to acquire 1D spectra from 2D data is being studied; while the next stage in the automation process, namely, how to extract useful information from the reduced 1D data, is the subject of our present study. Having considered the variety of different types of galaxy spectra, we find that there would be much to be gained if automatic separation of ELGs from non-ELGs is made in a new way. The ELG spectra tend to be highly redshifted, which often makes the identification of lines in them difficult. We find that a good criterion for the separating non-ELG and ELG spectra would be to test for the presence of a strong 4000 Å break in the continuum. This is because non-ELG spectra are generally at lower redshifts than ELG ones and tend to have large, easily-identifiable 4000 Å breaks. By contrast, the 4000 Å break is generally weak in ELG spectra.

The definition of continuum, even for physical spectra observed without any distortion, is not a trivial task. For galaxies, only the so-called "pseudo continuum" can be considered (Pelt 1990). The shape of the continuum is often determined by some low-degree polynomial fitting or by a broad bandpass filter. In these methods, the information in the vicinity of spectral lines and breaks is not used, and the fit is limited to spectral regions away from lines and breaks. Starck et al. (1997) have used à wavelet iterative methods to determine the local continuum and got good results. However, when automatic iteration is employed, problems will arise from regions of strong lines and breaks. For stars only, Bailer-Jones et al. (1998) used a median-filtering method that subtracts regions known to be affected by strong stellar lines. This method cannot be used for galaxy spectra though, because of the redshift. It also requires prior knowledge of each line and break.

In Section 2, we describe our procedure for estimating the positions and strengths of the 4000 Å break in galaxy spectra and for separating non-ELGs from ELGs. In Section 3, we describe our method of estimating the continuum level based on wavelet transforms. In Section 4, we measure the wavelengths of spectral lines and derive redshifts for non-ELGs. Finally, in Section 5, we apply our automated method to a sample of galaxy spectra and compare our results with those obtained by hand using IRAF.

2 THE 4000 Å-BREAK AND THE ELG/NON-ELG CLASSIFICATION

There are discontinuities in astronomical spectral continua. Most of these discontinuities are due to ionization. When the electron energy above an ionization limit is effectively unquantized, absorption bands are produced. For example, the ionization of hydrogen atoms causes the Balmer break at 3647 Å, while a large number of spectral lines of ionized metals, such as CaII H & K, produce the opacity that causes the 4000 Å break. In low dispersion spectra, the Balmer break is hidden in this opacity. Four spectral discontinuities, occurring at 2420, 2640, 2900 and 4000 Å, were considered by Bruzual (1983).

Zaritsky et al. (1995) used the 4000 Å break as a criterion in the classification of galaxies. The 4000 Å break is often used to determine the star formation characteristics of distant field and cluster galaxies (Poggianti et al. 1997). Tresse et al. (1999) studied the 4000 Å break of 1671 galaxy spectra and investigated the size of the break in both non-ELG and ELG galaxies.

The first step in our reduction procedure is to obtain, for each galaxy spectrum, the position of the 4000 Å break. In order to take redshift into account, we use a window covering the wavelength range 3950–4450 Å. This involved first searching for the largest discontinuity within the window. In practice this meant comparing the sizes of the discontinuities measured with respect to each bin of data within the wavelength range, and choosing the wavelength of the bin with the largest discontinuity. Our working definition of discontinuity strength, D, is taken to be

$$D_{4000} = \frac{\sum_{\lambda_D}^{\lambda_D+\Delta\lambda} f(\lambda)}{\sum_{\lambda_D-\Delta\lambda}^{\lambda_D} f(\lambda)}\,, \tag{1}$$

where f is the flux and λ the wavelength. Spectra with D_{4000} greater than 1.6 were attributed to non-ELGs and the rest to ELGs. This threshold was based on the results of Tresse et al. (1999). Bruzual (1983) and Tresse et al.(1999) used $\Delta\lambda = 200$ Å when computing D_{4000}. However, we used $\Delta\lambda = 50$ Å to find the position of the break even though we also used $\Delta\lambda = 200$ Å to compute D_{4000}. Because we did not know the position of the break exactly, using 50 Å in the small 3950–4450 Å range turned out to be more suitable. Based on trial and error, we found that it is better to use $\Delta\lambda = 50$ Å to find the position, but to use $\Delta\lambda = 200$ Å to judge between ELGs and non-ELGs.

Our 3950–4450 Å window only takes into account galaxies of $z < 0.125$. For $z > 0.5$, we have the additional problem that the 2900 Å break (and possiblly other breaks too) will enter this window. After the continuum estimation stage (see Section 3), our program therefore checks the number of absorption and emission lines against the ELG/non-ELG classification based on the 4000 Å break criterion. If a discrepancy occurs, then the continuum estimation step is repeated.

3 CONTINUUM ESTIMATION USING WAVELET TRANSFORMS

Although a filter bank can separate spectral lines from the continuum, it is difficult to fit spectral discontinuities, especially in the case of non-ELGs. This problem is apparent from Fig. 1a. In spite of this, it is important to fit the 4000 Å break as part of the continuum even if the galaxy concerned is a non-ELG. In order to achieve this, we generated pseudo-continua for non-ELG spectra by multiplying the flux levels by D_{4000} for all wavelengths shortwards of the observed wavelength of the 4000 Å break. This yielded pseudo-continua of the form shown in Fig. 1c. For the purposes of continuum fitting we worked with the pseudo-continua when dealing with non-ELG spectra.

The Fourier transform of a general 1-D distribution $s(v)$ in v space, can be replaced by a wavelet transform applicable to a small range, or 'window', in v. A wavelet transform consists of a family of waves generated by translations and dilations of $s(v)$ within the window concerned. It takes two arguments: v and s.

A wavelet transform is defined as

$$Wf(v,s) = \langle f, \psi_{v,s}\rangle = \int_{\infty}^{-\infty} f(\lambda)\frac{1}{B}\sqrt{s}\psi^*\Big(\frac{\lambda - v}{s}\Big)\mathrm{d}v, \tag{2}$$

where $f(\lambda)$ is a power spectrum; the base atom, ψ, is a zero average function, which centers on zero with a finite energy. ψ^* is the conjugate of ψ. A family of vectors is obtained by translations and dilatations of the base atom

$$\psi_{v,s}(\lambda) = \frac{1}{\sqrt{s}}\psi\Big(\frac{\lambda - v}{s}\Big). \tag{3}$$

This function is centered around v, like the windowed Fourier atom. In order to enable it to be applied to real 'histogram-type' data, a discrete form of the atom can be defined of the form:

$$\psi_{J,K}(\lambda) \equiv 2^{-J/2}\psi(2^{-J}\lambda - K), \qquad J, K \in Z \tag{4}$$

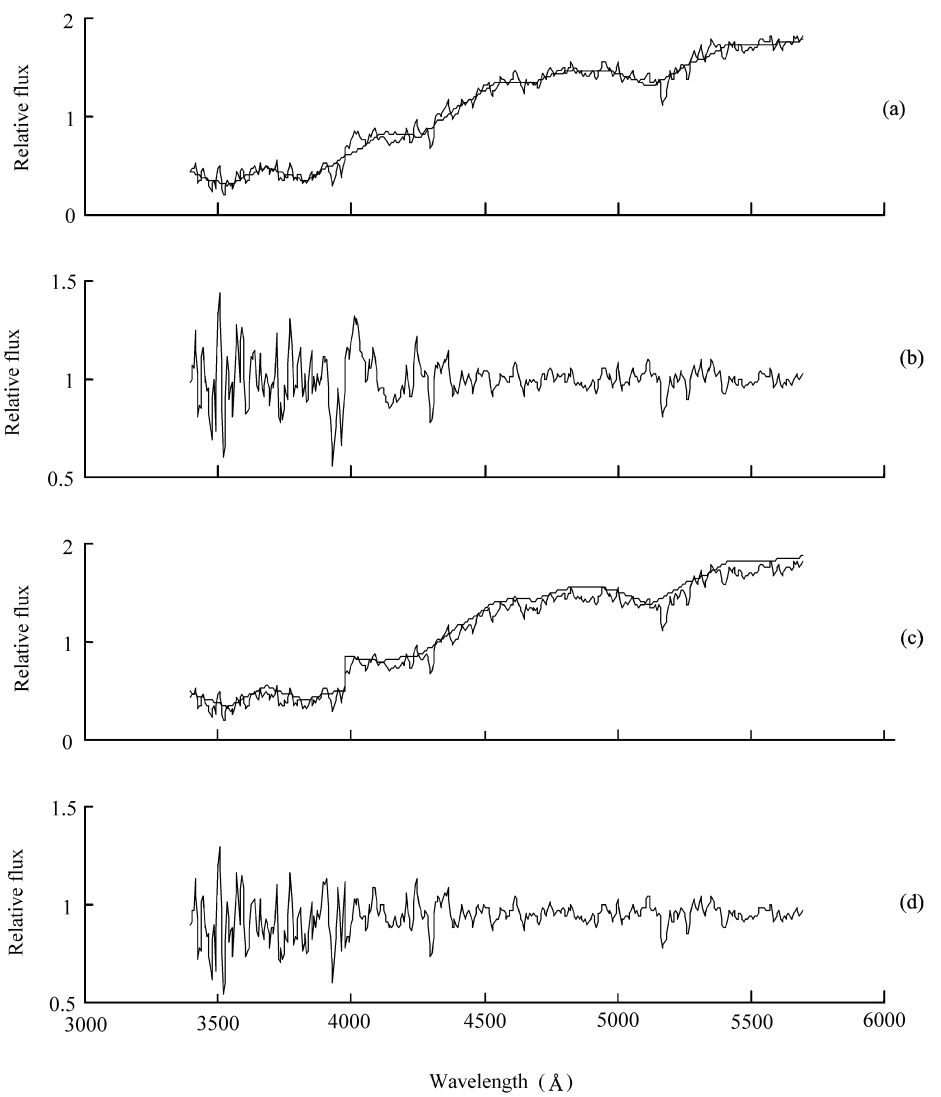

Fig. 1 An example fit to a 4000 Å break

where J is an integer representing a coarse measure of v, whilst K is an integer representing a coarse measure of the center of the transform. Unfortunately, it is difficult to compute the transform directly. Mallat (1989) developed a fast algorithm based on the concept of a multi-resolution analysis, in which a power spectrum $f(\lambda)$ can be represented by shifted and dilated versions of the prototype band-pass wavelet function ψ in Equation (3), and by shifted versions of a low-pass scaling function ϕ:

$$\phi_{J_0,K}(\lambda) \equiv 2^{-J_0/2}\phi(2^{-J_0}\lambda - K). \qquad J, K \in Z \tag{5}$$

Thus,

$$f(\lambda) = \sum_K u_K \phi_{J_0,K}(\lambda) + \sum_{J=-\infty}^{J_0} \sum_{J,K} w_{J,K}\psi_{J,K}(\lambda), \tag{6}$$

with $w_{J,K} \equiv \int f(\lambda)\psi^*_{J,K}(\lambda)d\lambda$, $u_K \equiv \int f(\lambda)\phi^*_{J_0,K}(\lambda)d\lambda$. It is therefore possible to reconstruct the signal from the wavelet coefficient $w_{J,K}$ and scaling coefficient u_K.

Wavelet transforms using a variety of base sets have been used in the analysis of astrophysical data for a number of years now. Of the several functions that have been invoked as wavelets to date, the set developed by Daubechies (1992) has proven to be particularly useful. This set was invoked by Meiksin (2000), in his analysis of quasar spectra because its component functions fall off rapidly with λ. In this paper, we have also chosen to use Daubechies's wavelets for the same reasons.

Because ϕ in Equation (5) is a low-pass function, the scaling coefficients u_K represent the shape of the continuum or of the pseudo-continuum $C(\lambda)$:

$$C(\lambda) = \Big(\sum_k u_k \phi_{J_0,K}(\lambda)\Big) d(\lambda) = C'(\lambda)d(\lambda), \tag{7}$$

where $d(\lambda)$ contains some information about the pseudo-continuum. The problem then becomes how to choose suitable values of J_0 (the number of J values in the set) and $d(\lambda)$ for real spectra. In the case of J_0 we first need to know the widths of the widest lines (emission lines for ELGs and absorption lines for non-ELGs) and/or bands in typical galaxy spectra. The greater the maximum width is, the larger J_0 needs to be, i.e. the larger the set of functions $J = 1, 2, 3, ...n, ...J_0$ needs to be. From an examination of the ELG spectral data of McQuade et al. (1995), Storchi-Bergmann (1995), and tables 4 and 5 of Jansen et al. (2000), it is found that the widest broad emission lines have FWHM of about 200 to 250 Å. As absorption lines tend to be narrower even in strong absorption-line spectra, we can therefore take the maximum width of any possible line to be 250 Å. In the wavelet domain, such a wide spectral line can only be detected when $J_0 \geq 6$ for a resolution of 10 Å per pixel or $J_0 \geq 7$ for a resolution of 5 Å per pixel.

Our program chooses J_0 taking into account the above finding, the resolution of the spectrum to be reduced and the type of galaxy spectrum (i.e. ELG or non-ELG). For an ELG spectrum of 10 Å resolution, this means starting with $J_0 = 6$ whilst for an ELG spectrum of 5 Å resolution, it means starting with $J_0 = 7$. For a non-ELG spectrum of 10 Å resolution, we start with $J_0 = 4$, whilst for a non-ELG spectrum of 5 Å resolution, we start with $J_0 = 5$.

The program then measures the widths of the widest lines in the spectrum concerned and reduces the J_0 values if the widest lines are less than 250 Å wide. It then computes the $C'(\lambda)$ of the spectrum, on the basis of the final J_0 value adopted. When J_0 has been chosen, a suitable

value of $d(\lambda)$ needs to be determined. Starck et al. (1997) used only the highest-order wavelet and an iterative algorithm to derive $C(\lambda)$. Because the continuum details are hidden amongst wavelets of different orders, we use wavelets of five different orders to subtract lines that have different widths.

$$d(\lambda) = \prod_{n=1}^{5} d_n(\lambda), \tag{8}$$

most of the information on d_n is contained in wavelet scales $(J_0 - 1)$ and $(J_0 - 2)$. Our program computes $C'(\lambda)$ first, and obtains $f_1(\lambda) = f(\lambda)/C'(\lambda)$, its low frequency component in the $(J_0 - 1)$th order wavelet domain which we shall call $d_1(\lambda)$. Then we obtain $d_2(\lambda)$ from the transform $f_2 = f_1(\lambda)/d(\lambda)$ in the $(J_0 - 2)$ scale. The steps are repeated four times in the $J_0 - n(n = 2, ..., 5)$ scale, yielding $d_3(\lambda)$, $d_4(\lambda)$, $d_5(\lambda)$. Figure 2 gives an example of how $d(\lambda)$ is calculated from $d_n(\lambda)$. This process eventually yields the continuum level $C(\lambda)$.

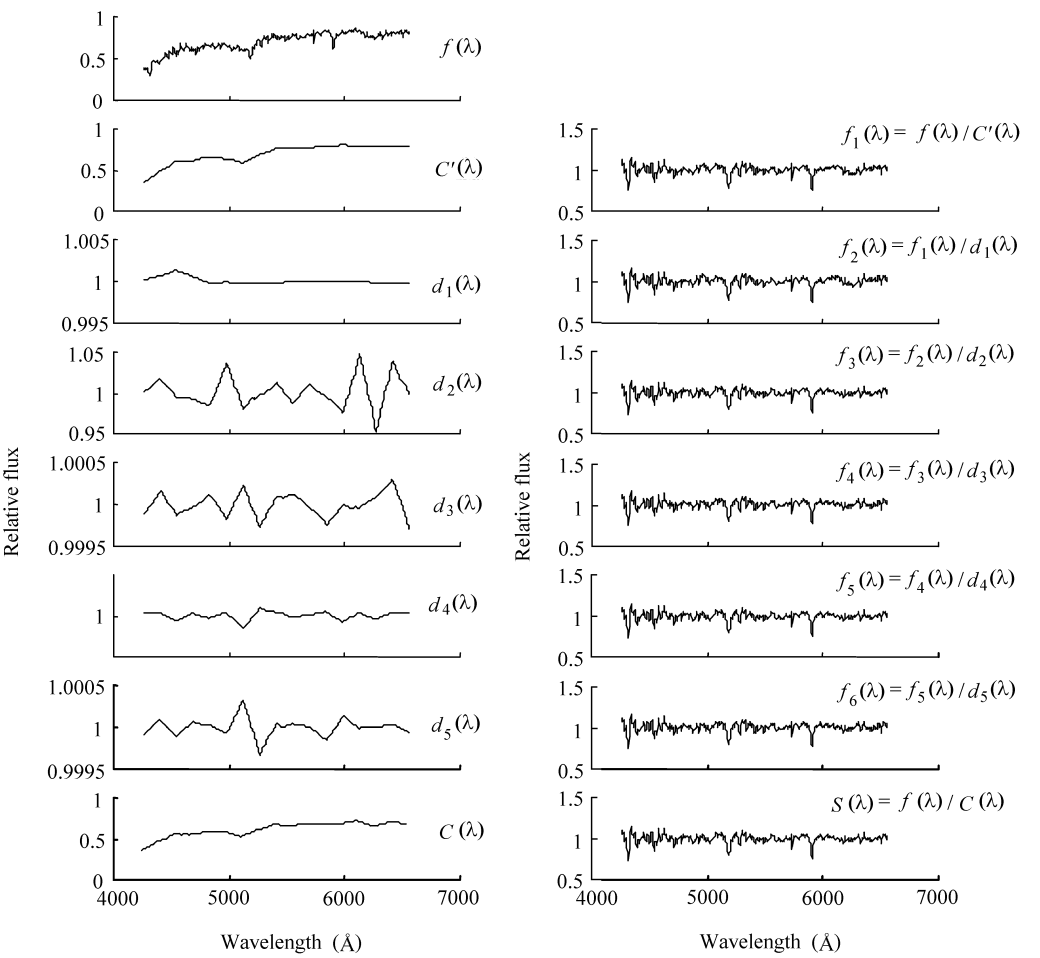

Fig. 2 An example of the procedure for separating the continuum from spectral lines

4 POSITIONS AND STRENGTHS OF SPECTRAL LINES

Stellar spectra tend to be dominated by a few strong absorption lines which change slowly across the spectral classes. A galaxy spectrum by contrast is more complicated as it is a combination of stellar spectra and emission lines. The simplest way to pick out strong lines in a spectrum with its continuum already subtracted, is to set a threshold. If the intensity of a feature exceeds the threshold, it is taken to be a line, otherwise it is treated as noise. In general, however, the noise varies with wavelength (Starck et al. 1997). Local noise thresholds are therefore needed when searching for genuine lines.

Our program divides each spectrum into many, separate, 10-pixel-long sections in wavelength space. The noise in each section is regarded as Gaussian white noise and the root-mean-square deviation (σ) is computed. We find that setting each local threshold to 3σ and fitting polynomials to both the upper and lower 3σ-limits ensure that only genuine lines are identified. The ideal section lengths (of 10 pixels) and clipping threshold (of 3σ) were obtained only after a great deal of trial and error. Sample polynomial fits to real data are shown in Fig. 3.

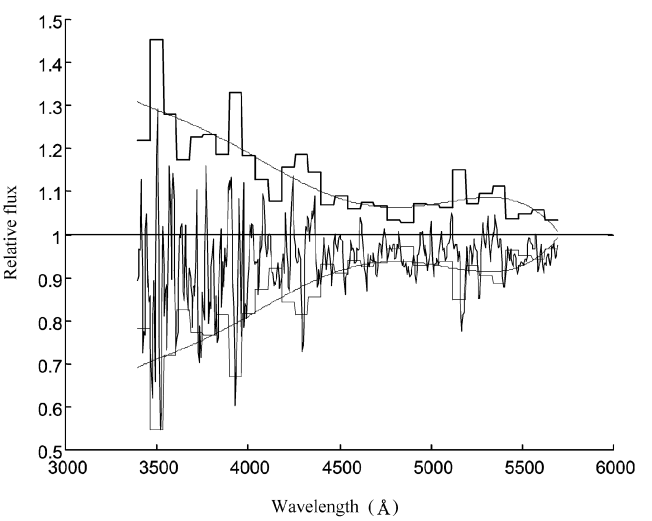

Fig. 3 An example of 3σ-clipping using local thresholds

After noise peaks and troughs had been excluded, the program fits the remaining strong lines with Gaussian models yielding the centroid wavelengths of the lines. If a spectrum is a non-ELG one, our program uses the 4000 Å break to compute the redshift and match the line-centroid wavelengths with the rest-wavelengths of known lines. The EWs of absorption lines are also computed by the program. For ELGs, on the other hand, only the peak heights (intensities) of emission lines (hence line ratios) are computed. How to identify these lines automatically by means of their line ratios is being studied.

The centroid wavelengths of all strong lines (both absorption and emission) are computed by our program according to the same formula as used by IRAF, namely:

$$\lambda_{\rm centroid} = \frac{\sum(\lambda(i) * (I(i) - C(i)))^{3/2}}{\sum(I(i) - C(i))^{3/2}}\,, \tag{9}$$

where $I(i)$ is the observed value of pixel i. ELG emission-line intensities are computed according to

$$\text{Intensity} = \frac{\sum((I(i) - C(i)) * \lambda(i))}{\sum \lambda(i)} \,. \tag{10}$$

The EWs for non-ELG absorption lines, on the other hand, are computed according to Pickles's (1998) formula:

$$EW = \Delta\lambda(1 - 2F_{\rm line}/(F_{\rm cb} + F_{\rm cr}))\,, \tag{11}$$

where $F_{\rm line}$, $F_{\rm cb}$ and $F_{\rm cr}$ are the average fluxes per unit wavelength within the line region and the pseudo-continuum regions lying to the blue and red of the line, respectively, and $\Delta\lambda$ is the width of the line region in Å.

5 TESTS ON A SAMPLE OF GALAXY SPECTRA

To test our methods, we use three sets of data. The first set contains the 55 galaxy spectra observed by Kennicutt (1992) using the Steward 2.3-m telescope. The spectral range covered was 3560–7100 Å with a resolution of 5–8 Å. The flux was normalized to that at 5000 Å, i.e., $F(\lambda)/F(5000)$.

The second dataset is of 31 spectra taken from McQuade et al.(1995). These spectra were obtained at KPNO with the 0.9-m telescope using the IRS. The wavelength range covered was 3500–8000 Å with a resolution of 10 Å. As there are seven galaxies in common with the first dataset, we omitted them and used only the spectra of the remaining 24 objects.

The third dataset was that obtained by Storchi-Bergmann et al. (1995) using the CTIO 1-m and 1.5-m telescopes. It has 48 galaxies. The wavelength range covered was 3200–10000 Å with a resolution of 5–8 Å. As there are 24 galaxies in common with the first two datasets, we use only those spectra of the 24 objects unique to this dataset.

When our program is used to reduce all 103 spectra from the three datasets described above, we find that four objects were classified erroneously as non-ELGs due to their large 4000 Å breaks even though they are really ELGs with relatively weak emission lines. These objects are NGC 1357, NGC 3147, NGC 5195 and NGC 6634. It should be possible to overcome this problem by using line strength information. However, this would require that individual lines be identified first. This is not all that straightforward as ELGs tend to have large redshifts. Nevertheless, we are now working on a more robust automatic spectral-line identification method, and this will be given in a subsequent paper.

In order to test the emission-line intensities, absorption-line EWs and line-center positions derived by means of our program, we use the 'splot' package of IRAF to determine these values by hand. The values obtained using our program and those obtained using IRAF are compared in Fig. 4. Figure 4a shows the differences found in ELG emission-line intensities: in 97 cases out of a total of 431, the difference is greater than 20% but in no case does the difference exceed 40%, whilst Fig. 4b shows the differences found in absorption-line EWs for non-ELGs: in 132 cases out of a total 1019, the difference is greater than 15% but in no case does the difference exceed 20%. These results show that our program is precise enough to obtain intensity information from spectra. Figure 5 shows the differences in the line-centers obtained by the two methods. Most of the discrepancies in the centroid-wavelength values are smaller than 10 Å, i.e. one pixel. They are caused by limitations in spectral resolution. For non-ELGs, we also compare

the redshifts obtained from measurements of the positions of the 4000 Å breaks, as shown in Fig. 6.

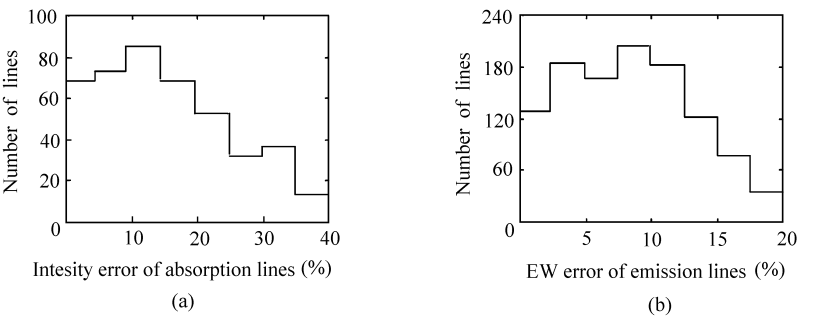

Fig. 4 (a) Comparision between absorption-line intensity obtained using our program and those obtained using IRAF; (b) Comparision between emission-line EW values obtained using our program and those obtained using IRAF.

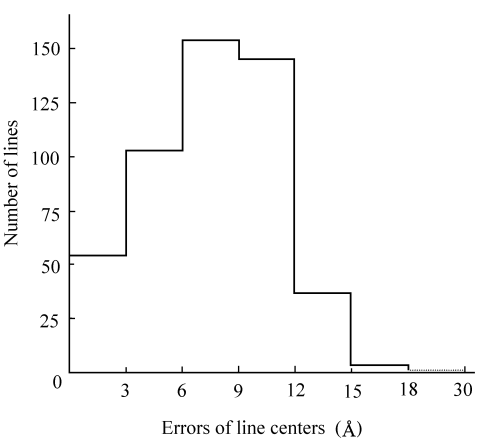

Fig. 5 Comparison between centroid-wavelength values obtained using our program and those obtained using IRAF

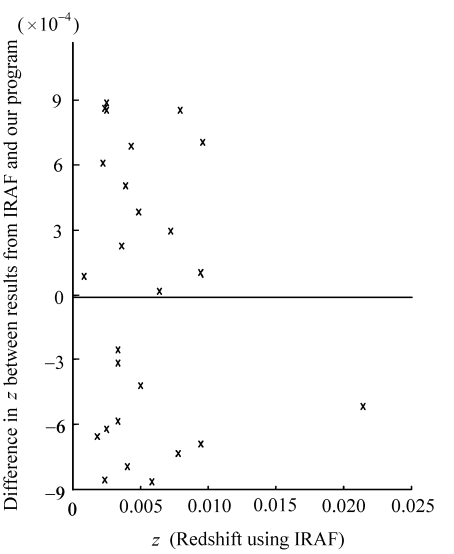

Fig. 6 Comparison between non-ELG redshift values obtained using our program and those obtained using IRAF

6 DISCUSSION

Zaritsky et al. (1995) have shown that strong lines and the 4000 Å break contain sufficient information for an automated spectral classification of galaxies into broad types, and that in order to obtain pure sets of strong lines without any false line one must make a good fit to the pseudo-continuum. The continua around the discontinuities determine directly whether the false lines will be "detected" or not. Because non-ELGs have high values of D_{4000}, unless the

4000 Å break is modeled properly, false lines will be "detected". We obtain uncontaminated sets of strong lines (i.e. without any false line in them) by setting a reasonably high threshold and presumably because there are no significant problems with our pseudo-continuum estimation procedures.

There are several breaks in any galaxy spectrum. Because the redshift of all of the non-ELGs we used in this paper are relatively small, only their 4000 Å breaks lie within our observed optical wavelength ranges. In the case of ELGs, however, the 4000 Å breaks tend to be small and cannot generally yield much useful information. However, if the redshift is large enough, other breaks, such as L_α may lie within the optical wavelength range of our spectra and be detectable by our program. Although our program can compute the relative intensities of emission lines, how best to measure redshifts and identify individual emission lines in optical ELG spectra of widely varying redshift will be the subject of a future paper.

In the cases of non-ELG spectra, we have demonstrated that our program successfully uses the position of the 4000 Å break to identify strong absorption lines, and that once these lines have been identified it computes reliable redshifts for them.

Although we cannot yet compare our method with the results obtained by the SDSS and 2dF galaxy surveys, because their spectral data are not publicly available yet, we have obtained consistent results from three datasets based on observations made by other authors. Further steps towards a fully automated classification pipeline for LAMOST spectra will be presented in forthcoming papers.

Acknowledgements We would like to thank Jingyao Hu and Huoming Shi for useful discussions as well as Ke-shih Young for much work on editing the manuscript. The authors gratefully acknowledge support from the LAMOST project.

References

Bailer-Jones C. A. L., Irwin M., von Hippel T., 1998, MNRAS, 298, 1061
Bruzual A. G., 1983, ApJ, 273, 105
Daubechies I., 1992, Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics, University of Lowell, Mass. June 1990, Philadelphia: Society for Industrial and Applied Mathematics (SIAM)
Jansen R. A., Fabricant D., Franx M. et al., 2000, ApJS, 126, 331
Kennicutt R. C., 1992, ApJS, 79, 255
Lewis I. J., Glazebrook K., Taylor K., 1998, In: S. D'Odorico ed., Optical Astronomical Instrumentation, Proc. SPIE, 3355, 828
Mallatt S. G., 1989, IEEE Transactions on Patten Analysis and Machine Intelligence, 11, 674
McQuade K., Calzetti D., Kinney A. L., 1995, ApJS, 97, 331
Meiksin A., 2000, MNRAS, 314, 566
Pelt J., 1990, Bull. Inform. CDS, 38, 95
Pickles A. J., 1998, PASP, 110, 863
Poggianti B. M., Barbaro G., 1997, A&A, 325, 1025
Starck J. L., Siebenmorgen R., Gredel R., 1997, ApJ, 482, 1011
Storchi-Bergmann T., Kinney A. L., Chillis P., 1995, ApJS, 98, 103
Tresse L., Maoddox S., Singleton C., 1999, MNRAS, 310, 262
Wang S., Su D., Chu Y. et al., 1996, Applied Optics, 35, 25
Watson F. G., Parker Q. A., Bogatu G. et al., 2000, In: M. Iye, A. F. Moorwood, eds., Optical and IR Telescope Instrumentation and Detectors, Proc. SPIE, 4008, 123,
York D. G., 2000, AJ, 120, 1579
Zaritsky D., Zabludoff A. I., Willick J. A., 1995, AJ, 110, 1602