Bull. Spec. Astrophys. Obs., 2003, 56, 126-133

Methods for graphic data input from paper medium

V.V.Grechnev[a], V.E. Abramov-Maksimov[b], N.G. Peterova[c], T.P. Borisevich[b], T.I. Kaltman[c], N.S. Meshalkina[a]

[a] Institute of Solar-Terrestrial Physics of SD RAS, Irkutsk, Russia
[b] Central astronomical observatory at Pulkovo, Saint-Petersburg, Pulkovo, 196140, Russia
[c] Special Astrophysical Observatory of the Russian AS, Nizhnij Arkhyz 369167, Russia

Received July 18, 2003; accepted July 28, 2003.

Abstract. Valuable large arrays of data recorded on paper medium have been acquired to the present time in various areas of knowledge. The bulk of these data has not yet been analyzed. Entering the data in computers promises their digital processing and analysis taking advantage of powerful mathematical tools and computing facilities. However, the digitization of paper records is a complex problem ranking among tasks of pattern recognition. We address here simple ways to solve this problem. Having divided the task into several subtasks, we can solve some of them automatically using image processing methods and statistical analysis. Next, we commit the procedures which we cannot describe formally to the operator who specifies required actions by means of the graphics user interface. We demonstrate our methods with digitizing a paper record of a solar observation made with the Large Pulkovo Radio Telescope.

Key words: techniques: image processing – methods: data analysis

1. Introduction

Valuable large arrays of data recorded on paper medium have been acquired to the present time in various areas of knowledge. The bulk of these data has not yet been comprehensively analyzed. Such arrays are available practically in any astronomical observatory. An example is the archives of solar microwave emission observations with the Large Pulkovo Radio Telescope (BPR) (Khaikin et al. 1964). The archives contain the data to embrace three solar cycles (Abramov-Maksimov et al. 1999). These data is a uniform series obtained with unified observations and processing routines. Using the data, one can perform statistical analyses based on an extensive experimental material and study consistent patterns that are manifest at time scales of order of 11-year solar cycle. Some portions of the archives of solar observations with the radio telescopes RATAN-600 and SSRT are still on paper medium. Similar tasks also arise if, e.g., some graphs to be analyzed are presented in papers published in the previous years, etc.

Transformation of graphic data from paper medium into the digital form is really important because their input into computers would allow digital processing and analyses the data taking advantage of powerful mathematical methods and computing facilities. However, the digitization of paper records is a complex problem ranking among tasks of pattern recognition. Nevertheless, it is possible to find rather simple solutions for transformation of graphic information from paper medium into digital arrays. By dividing the task into several subtasks, we can solve some of them automatically in a relatively simple way. Next, we commit to the operator the procedures which we cannot describe formally. He or she makes decisions and specifies in the interactive mode actions to be done. In particular, the problem of pattern recognition is solved by the operator by means of the graphics user interface (GUI).

2. Stages of record digitization

We distinguish the following steps in the digitization of graphic records:

- preparation of the material for scanning;
- scanning;
- preparation for transformation to a digital array;
- transformation itself, i.e. digitizing the image;
- editing (retouch);
- calibration;
- combining portions, if the whole record cannot be scanned in a single pass;
- straightening image.

It is important to provide for the calibration and combining techniques before the scanning to prevent irreparable losses of information. It is also important to retain the scale and coordinate system of the

scanned image during the whole subsequent digitization process.

2.1. Preparation for scanning

Since long records on paper medium (usually tape) generally exceed the format which can be scanned in a single pass, they have to be broken up into portions ('frames') to be scanned separately. The frames basically can be combined provided that each frame contains at least two tick marks of time. However, the recorder speed can be high, and in such a case the distance between the tick marks exceeds the horizontal size of the frame. Besides, the tick marks of time may be absent in some of the records for technical reasons. Therefore, to ensure subsequent combining the scanned frames, one may put additional marks at the edges of the frames manually. Each of them must be present in both adjacent frames.

2.2. Scanning

When entering data from paper into a graphics file by means of a scanner, it is important to ensure:

- saving all information, i.e., no losses of resolution and color/halftone information. There is no need to keep the 3-byte True Color mode; the scanned images, however, must not be visually worse than the original image. One byte per pixel is sufficient, i.e., the halftone mode for monochrome images, and the Pseudo Color mode with a color table for color images. The resolution must be sufficient to resolve both graph and reference grid lines, i.e., to represent the thickness of these lines by a few pixels;
- homogeneity of data, i.e., the same scale, color table, and file formats for all the records.

As we already noted, the means to combine the frames into a whole record as well as the calibration technique for both coordinates should be developed beforehand. Otherwise, the scanned files may not be processed completely.

An example is presented in Fig. 1 which shows a scanned portion of the total intensity record of the solar radio emission made with the BPR on 2 September 1971 at a wavelength of 3.2 cm. The color image in the figure is schematically shown by gray halftones. In reality, the upper curve is red, and the rest of them are gray.

2.3. Straightening the image

When the operator works with a large number of paper records, he (she) may likely place the paper on the scanner glass not strictly in parallel with the direction of scanning, but slightly turned. One can try to find the turn angle automatically using the following fact. If we compute the sum of the image elements over one of coordinates, then the reference grid of the other dimension appears as dashes on the resulting line. The modulation depth (the relative height of the dashes) has a maximum if the orientation of the image is strictly in parallel to the grid. Performing a series of turns of the image by small angles and monitoring the modulation depth, one can find the proper turn angle corresponding to the correct orientation of the reference grid with respect to the direction of scanning. The result is checked by the operator by means of the GUI. If it is not satisfactory, the operator marks four check points in the corners of the image to specify more accurately the angle of turn, and the turn correction is performed referring to these check points. This procedure may be performed prior to the digitization of the image or in the course of the calibration.

2.4. Preparation for the transformation to a digital array

This stage is necessary to isolate in the image the points to be digitized, and to remove all others (spots, reference grid, lines of other graphs, notes and other inscriptions, etc.) Attempts to solve this task automatically meet the greatest difficulties, because this is virtually extraction of the useful signal at a signal-to-noise ratio of unity or even less. Here, the noise signal is not quite regular due to various-scale inhomogeneities of the background, brightness, contrast, and thickness of the grid as well as other lines in the image, etc. This is why the preparation stage is accomplished in several steps.

Hereafter we call the curve to be digitized the *graph* for shortness.

Filtering. It is usually not possible to suppress noises efficiently in an automatic way because of the various-scale inhomogeneities of irrelevant lines (grid etc.) and close thickness of the graph and grid lines. Therefore, filtering is used to perform the amplitude selection of the pixels belonging to the graph and noise. To suppress the grid, the Fourier filtering can be applied. Besides, the grid lines are well-defined on the one-dimensional vertical and horizontal sums.

Selection of regions to be digitized. If the graph line is well pronounced owing to its color or brightness, it can be effectively distinguished by analyzing the distribution of the pixel values. Just this method is used in processing the BPR records. In this case, regions of the increased population are present in the histogram representing the density function of the amplitude distribution of pixels. In the BPR records, those regions are discriminated by a 10^{-4} level of the histogram maximum (i.e., approximately this fraction of the entire area of the image). Typically this results in 15 to 18 variants of the graph line selection. Thus Fig. 2 shows 16 variants for the

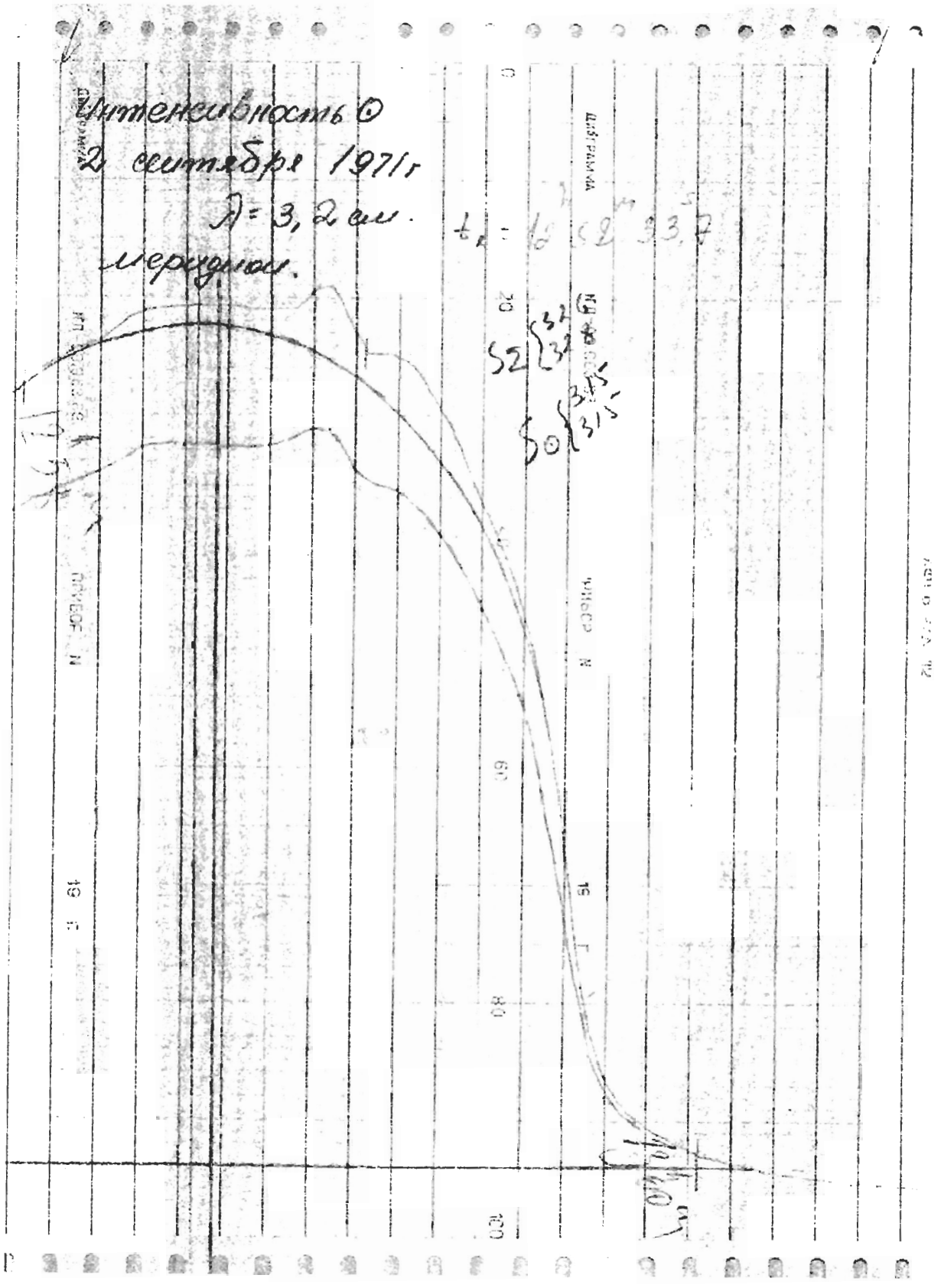

Figure 1: *Example of a scanned portion of the solar radio emission: one-dimensional intensity scan recorded with the BPR at* $\lambda = 3.2$ *cm on 02.09.1971.*

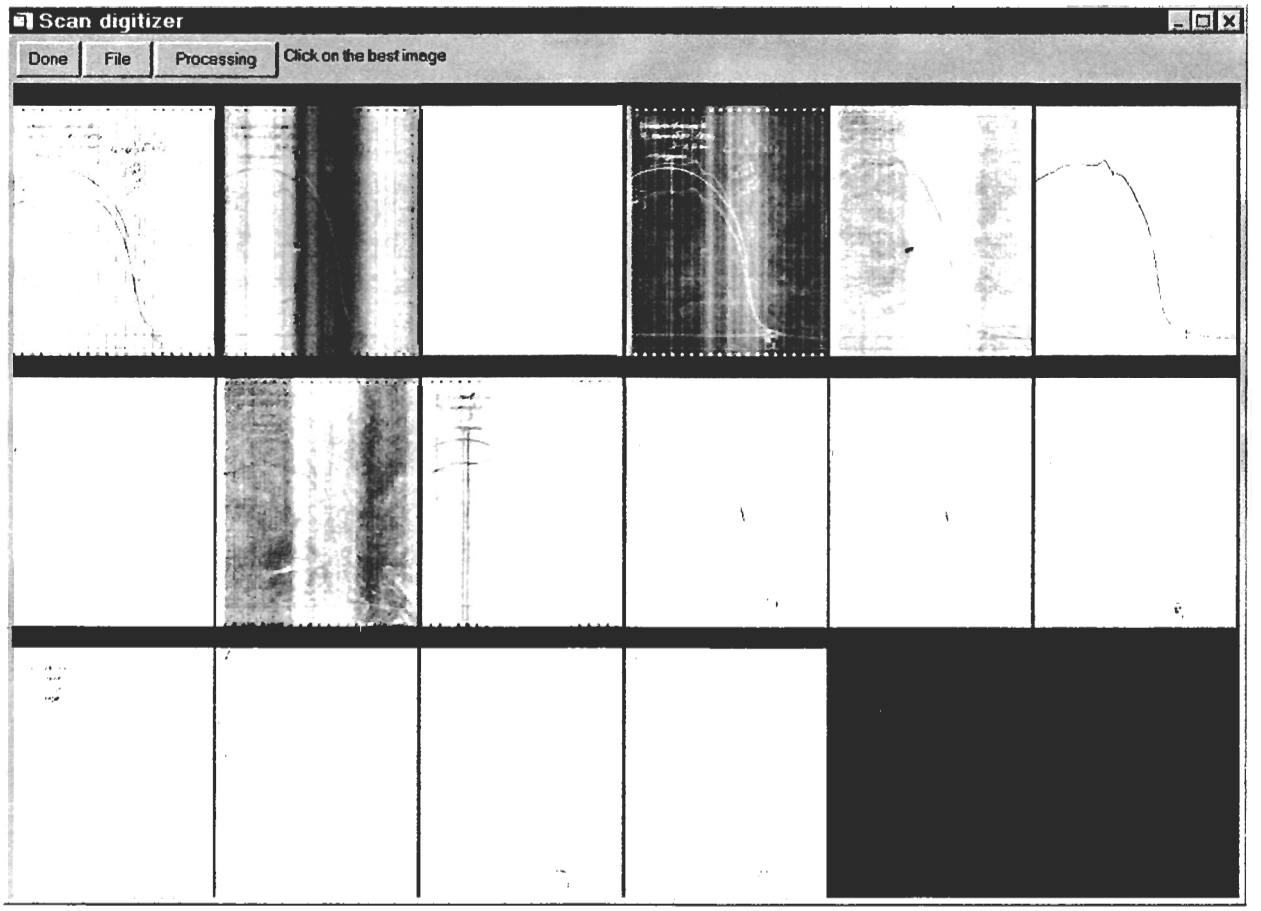

Figure 2: *Variants of amplitude filtering of the scanned fragment of the BPR record carried out at* $\lambda = 3.2\,cm$ *on 02.09.1971 (Fig. 1).*

portion of the record from Fig. 1.

All these versions displayed as a logical (monochrome) array are presented to the operator. In this array, the pixels whose values range within the limits found are shown in black (zero) against the white background (unity). The operator indicates the most appropriate of the presented versions of the amplitude filtering of the image, which is used to eliminate the grid lines.

This technique allows readjustment-free processing images of different color characteristics. It also allows distinguishing curves of different colors in one image.

Concatenation of graph segments. Preceding operations result in an approximate selection of the pixels belonging to the graph. At the same time, noisy spots are still present in the whole image, but their density outside the graph area is reduced. Therefore, they can be eliminated by means of a mask superimposed on the image. The mask should be thicker than the graph line and embrace the graph region.

The region of the graph is selected in the following way. First, a mask is formed using the morphologic dilation of the image by a disk structuring element (for details see Haralick et al. 1987). This results in blurring the lines and filling the gaps, and the mask becomes consisting of a number of rather large spots. Then a search is performed for all topologically linked regions (spots) in the masked image. Few regions associated with the graph obviously have large areas, whereas the areas of compact defects are small. For this reason, the analysis of few regions with the largest areas, i.e., of the highest population, is sufficient (7 in our case). These regions are presented to the operator, and he (she) specifies which of them should be considered in the digitization. The marked regions are merged, the rest of them are excluded, and the result is employed as the mask superimposed on the image, i.e., the mask and the image are multiplied. Irrelevant points and spots scattered over the field are thus eliminated, and the gaps in the record are filled.

Manual editing. When the image quality is high, and the number of irrelevant lines and defects is small (for instance, when the image is obtained by scanning a published diagram, and the graph is surrounded by coordinate axes only and has no grid), the simplest way of preparation of the image for digitization is

the manual retouch using a graphics editor. We have developed a GUI-supplied digitizer application with a built-in graphics editor for such simple processing. Note that, in addition to the standard modes of editing, erasing of the image *outside* the marked region appeared quite efficient.

2.5. Transformation itself (digitization of the image)

The transformation of the image to a digital array is performed by the search in the image for values different from the dominant value of the field. The imperfection of the images contained in the array (X, Y), as well as the finite line thickness, results in the fact that several values of Y generally correspond to each X value. To eliminate this ambiguity, we form another array (X^1, Y^1), in which there is no repetition of the X^1 values. Depending on the chosen way of transformation, the maximum, mean, or minimum value Y^1 of Y corresponds to each X^1:

$$Y_i^1 = \begin{cases} \min(Y_{ij}) \\ \frac{1}{n}\sum\limits_{j=1}^{n} Y_{ij} \\ \max(Y_{ij}) \end{cases}$$

The most typical is the choice by the mean value, that is, simple allowance for the line thickness. The choice by the maximum value is expedient if impulsive features in the record are of interest. The selection by the minimum value is convenient if the lower envelope of the record (trend) is important.

The obtained values of (X^1, Y^1) are interpolated to a regular grid (X^0, Y^0). Possible gaps between the points of the graph are filled herewith.

2.6. Editing (retouch)

The task of editing (retouch) is elimination of inevitable errors in automatic digitization. Besides, the presence in the records, e.g., insets of time tick marks also interferes with handling data that are already digitized. The graph can be smoothed in the course of retouch within the intervals containing such interferences. The simplest way of editing is the marking of the start and the end of each interval that requires retouching in the interactive mode (using GUI). The values within these intervals are replaced by the results of interpolation from the values at the ends of the intervals.

2.7. Calibration

We emphasize once again that the calibration technique should be elaborated prior to scanning, otherwise data already digitized may appear not possible to calibrate.

If the digitization of the image containing the coordinate axes or other similar means of calibration is executed, the calibration can then be performed referring to two points (X_{01}, Y_{01}) and (X_{02}, Y_{02}) whose corresponding values (x_{01}, y_{01}) and (x_{02}, y_{02}) are known. The calibration accuracy is better if those reference points are localized in the opposite parts of the image. The calibrated value is

$$X_c = \frac{X - X_{01}}{X_{02} - X_{01}}(x_{02} - x_{01}) + x_{01},$$

and the calibrated value of Y_c is calculated in a similar manner.

If a logarithmic scale is used in the given dimension, then, correspondingly,

$$X_c = 10^{\frac{X - X_{01}}{X_{02} - X_{01}}(\lg x_{02} - \lg x_{01}) + \lg x_{01}}.$$

However, the records often do not contain any reference means for the calibration of the Y values. In such a situation, the calibration is performed by means of other methods using the properties of the recorded signal itself, and in the course of the digitization process it is only required to maintain for all scanned portions the exact position of paper with respect to the edge of paper medium and the same scale of the graph. For instance, if source paper medium is a tape, then the height of both lower and upper edge lines of the tape should be registered.

2.8. Combining portions

To combine digitized portions of the graph, accurate calibration along the horizontal coordinate is necessary. It can be performed, for instance, if each portion contains at least two time tick marks, and their values are known. In this case, it is possible to calibrate all portions separately before combining, and no additional means for the concatenation are needed.

However, it is not always the case where the time values of all tick marks are known (see also section 3). In such cases, it is advisable to provide the record portions with markers. These markers must ensure the concatenation and calibration accuracy, i.e., either be thin and oriented parallel to the line of cutting the record into fragments, or unambiguously point to one of the vertical grid lines that contains paper medium (tape). When digitizing the BPR records, the combining is executed referring to such markers.

When combining, all the necessary parameters inscribed on paper medium are also entered (in the case of BPR records on tape, these are the date of observations, culmination time, wavelength, and the Stokes parameter recorded). These parameters are read out by the operator visually from the image and entered into the application which adds them programmatically to the header of the FITS file being formed (Fig. 3).

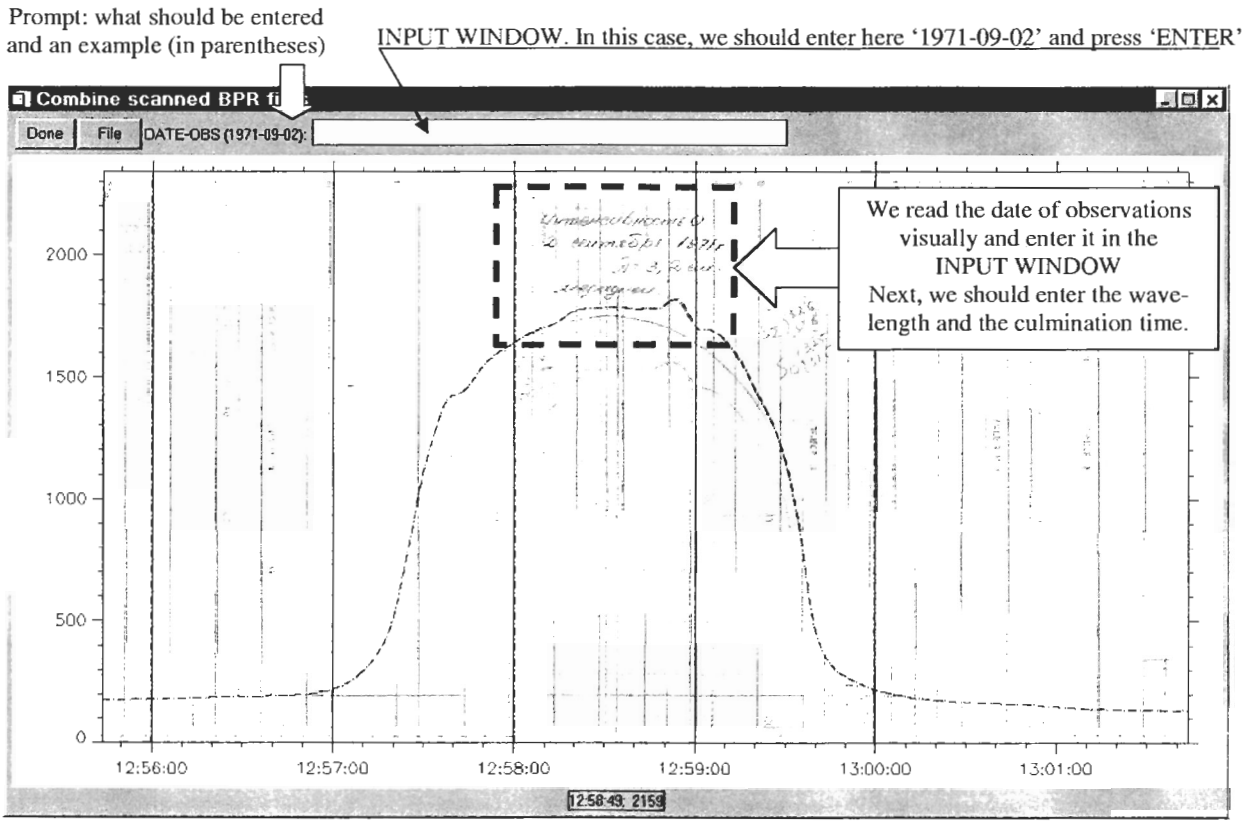

Figure 3: *A screen dump of the application for combining digitized portions.*

The values obtained are interpolated to a regular grid in both dimensions.

2.9. Interpolation

Interpolation is necessary for filling the gaps in digitizing and editing. The digitized record may subsequently undergo any transformations, in particular, differentiation. In this connection, it is attractive to demand continuity of the obtained function as well as its derivatives. Spline interpolation meets such a requirement, however, it can be artifactitious. Linear interpolation is free from this feature, but the differencing within the intervals interpolated produces horizontal bars. At the present time, we use in the application linear interpolation only. The points between which the data are reconstructed using interpolation are recorded in the header of the output FITS file, and the whole additional information entered in the course of processing is logged as well.

3. Conclusion

We developed techniques and applications implementing them to digitize graphical data recorded on paper medium. The applications perform preparation of the scanned image for transformation to a digital array, the transformation itself (digitization of the image), editing the curve obtained in the interactive graphics mode, calibration and combining data subsets digitized with interpolation of gaps, and referring them to a common, regular grid as well. If it is difficult to formulate a robust algorithm for solving a particular subtask, the decision is made by the operator who specifies the required actions using the GUI.

For implementation of our techniques, we employed the IDL programming language[1] currently extensively used by solar astronomers. IDL is convenient, in particular, owing to its high level, the presence of a rich set of procedures realizing various mathematical methods and images processing techniques, and accessibility of files of popular formats. The debugging of IDL codes is efficient thanks to the interactive mode, and the tested routines can be simply incorporated into the main GUI-supplied application. In particular, we used digitizing applications in studies published elsewhere (Altyntsev et al. 1998; Grechnev et al. 2001). To store intermediate and output data, we employ the widely used in the astronomy computer-independent FITS format which allows recording digital arrays of any accuracy along

[1] IDL is the trade mark of the Research Systems, Inc., http: www.rsinc.com

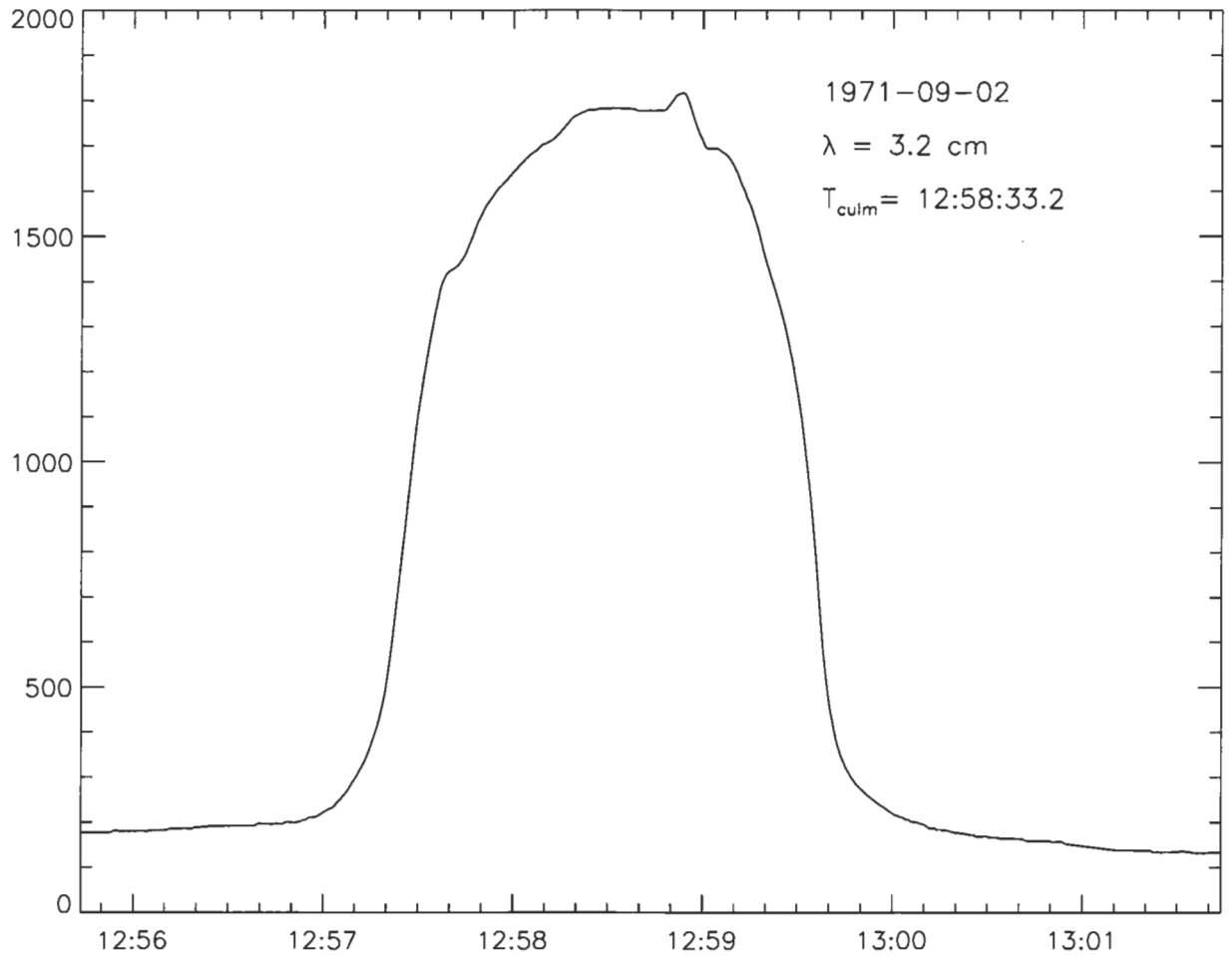

Figure 4: *Entire digitized solar scan observed with the BPR on 02.09.1971 at* $\lambda = 3.2\,cm$ *(3604 points).*

with accompanying parameters and comments.

Some preliminary stages of digitization and combining a record of solar observations with the BPR on 02.09.1971 at $\lambda = 3.2\,cm$ are shown above (see Fig. 1 and 2), and the final version of the entire scan is shown in Fig. 4. This data array recorded in the FITS format (typical length is 50 to 70 kbytes) is suitable for subsequent digital processing and analyses.

Thus we have achieved certain success in the transformation of paper graphical data to digital arrays. In particular, first promising results of application of the methods developed to digitizing large paper archives of one-dimensional solar observations at the BPR are obtained. Using these data, one can, e.g., study the periodicity of the solar activity, develop methods for the forecast of powerful flare events, and search for interrelations between the solar activity and the terrestrial phenomena, e.g., the variations of the weather on the Earth. All these problems are topical, and the interest in them is still growing. The shift of the BPR solar observation archives to the digital form would facilitate and expedite its processing and make it possible to continue the studies mentioned above (Korobchuk et al. 1984, Peterova et al. 2000) using the material of higher statistical importance. The availability of digital data also permits conducting new studies which were not possible when dealing with the paper archives because of the laboriousness of processing and insufficient measurement accuracy.

Acknowledgements. This work is supported by RFBR grants 02-02-16548, 03-02-16591, the program "Integration" (the project No. F0027/2308), SS-477.2003.2, the Russian Federal Program "Astronomy", and the program of SPbRC 2002 (the project "Upgrade of the BPR").

References

Abramov-Maksimov V.E., Bogod V.M., Borisevitch T.P., Korzhavin A.N., Opeikina L.V., Peterova N.G., 1999, Proc. of the VIII Russian-Finnish Symposium on Radioastronomy, June 28-July 3, Saint Petersburg, 108

Altyntsev A.T., Grechnev V.V., Hanaoka Y., 1998, Solar Physics, **178(1)**, 137

Grechnev V.V., Altyntsev A.T., Lesovoj S.V., Yan Y., Fu Q., 2001, Issled. po Geomagn., Aeronomii i Fizike Solntsa, Novosibirsk, Nauka, vyp. 113, 69

Haralick R.M., Sternberg B., Zhuang X., 1987, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-9, No. 4, 532

Khaikin S.Eh., Kajdanovskij N.L., Esepkina N.A., Shivris O.N., 1960, Izv. GAO USSR AS, **164**, 3

Korobchuk O.V., Peterova N.G., 1984, in: Transactions "Radioizluchenije Solntsa", Publ. LGU, L., vyp.5, 102

Peterova N.G., Shpital'naja A.A., Tsyrkunov V.S., 2000, in: Transactions "Fundamental'nyje problemy estestvoznanija", SPb, **II**, 96