Spectrum Digitizing and Conditioning for the Infrared Library of Amorphous Silicon Alloys

R. Esen* and H. Kavak

Physics Department, University of Cukurova, 01330 Adana, Turkey

Received April 14, 1994*

A simple method for digitizing of infrared (IR) spectrum from scanned images of published data is proposed and tested for amorphous and polycrystalline silicon alloys. The method uses the following special constraints of IR spectral data: (a) scanned image should have only two gray levels (black and white); (b) each point on the graph should be single valued; (c) frequency distribution of non-background pixels of columns should give a maximum, at the number of scan lines in the images. These criteria along with the closest distance approach are used in programs for identifying IR spectra and found to be successful in more than 95% of the cases. After identification of the lines, base line correction, normalization, and noise reduction techniques were applied before recording the data to the magnetic media. This method can be used for purposes other than forming an IR reference library, like archiving, regraphing, or comparing any spectral data.

1. INTRODUCTION

A complete IR reference library of amorphous and polycrystalline semiconductors does not exist. Since IR spectra of semiconducting alloys are dependent on the deposition technique used, one needs all the fabrication methods to get the whole library, a condition which is almost impossible to fulfill. Another difficulty in obtaining such a library is that in many cases required properties of target materials (mobility, resistivity, energy gap, etc.) can be obtained at nonstoichiometric ratios, thus resulting in a great number of spectral data series. For these reasons it is decided that, for accumulating an IR reference library, using the published data would be a convenient method. So, a method for digitizing the published spectral data is needed.

The digitization process can be performed by using either a conventional digitizer or a scanner and by processing with an appropriate software tailored for this purpose. Use of a digitizer in the digitization process has two disadvantages: the operator errors and the speed and precision losses. Since the operator has to decide the correct position of the tip of the digitizing head, clearly this predigitizing alignment can induce an operator error. The digitization speed is the other disadvantage of using a digitizer instead of scanning. For example, a full IR spectrum with four scan lines may require several hours to digitize while digitizing by scanning and its postprocessing may require less than 10 min with the presently proposed method and developed software. Also, the scanner is very common since it is used for other purposes. At the time of this writing there was no commercially available software designed especially for spectrum digitization from scanned graphical images.

Even though an IR spectrum is available in digital form, some sort of signal conditioning should be used in order to make it suitable for the library search and quantitative analysis software. The signal conditioning part of this work uses base line correction routines required by this special case.[1]

2. METHOD

2.1. Preconditioning of the Image. For the purpose of the present work, the images of the IR spectral data were taken with an A4 flatbed scanner model 216 at a resolution of 300 dpi (dots per inch). The scanner was controlled by a proprietary software from Apple with an Apple Macintosh Classic Computer Model LC.

Since it is common both for DOS and Macintosh environments, TIFF (tagged image file format) was used for the scanned images. Later, these scanner files were converted to DOS file formats for analysis with an IBM-PC compatible computer environment. Programs are coded in Pascal language.

To display the spectral images, pixels are forced to be white or black. The background color is taken as white and the lines are accepted as black as in the spectral data. Since there may be a lot of gray points in the scanned images (resulting from photocopying and/or scanning), these pixels should be classified to either white or black. A simple logic is used in this binary classification[2] step:

(a) If the pixel in question is isolated, then it belongs to the background.

(b) If the pixel is not isolated, then it is checked to determine if it is darker than the half of the gray level scale (256). If this condition is met, then the point is accepted as black; otherwise it is accepted as a white point.

2.2. Identifying the Spectra. For deciding how many scan lines are present in the data to be classified, data are read from the screen columnwise, and vertically connected points are represented by a single pixel at the average y value of those neighboring points. The reason for this transformation is to eliminate the complexities that may arise from nonhomogeneity of the line thickness. Later the frequency distribution of the columns is calculated. Frequency is defined as the number of nonwhite points in a column. Excluding the horizontal axes, the maximum frequency must give the number of scan lines. This assumption worked perfectly at all multiple and single scan spectra.

In the next step, all the columns were reevaluated according to their nonwhite point values. If this number is found to be equal to the number of scan lines in the spectrum (found from the frequency distribution peak value), these pixels are taken as candidates for the spectrum lines. Then, they are put in a matrix according their respective order in the column (i.e., y_{1i}, y_{2i}, ..., y_{ni} are accepted as ith columns, first row, second row, etc.). After the first pass of the classification process all suspicious points are excluded from the classified set. Exclusion criteria were as follows: if the candidate point is isolated, (i.e., it does not have at least three adjacent candidate points), the points in this column may not belong to the real scan lines.

* Abstract published in *Advance ACS Abstracts*, June 15, 1994.

0095-2338/94/1634-1158$04.50/0

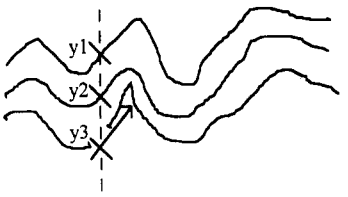

Figure 1. Broken lined spectrum showing the importance of the candidate validation process.

The noise, symbols, and lettering in the spectrum may give these effects. Sometimes, some of the candidate points may belong to the scan lines, but their respective order may be different, as the example in Figure 1 shows.

In this figure, there are three scan lines. The arrow coincides with the broken third line, so at that point one gets the number of scan lines as 3, which is correct; but y_1, y_2, and y_3 do not all belong to real scan lines: y_3 belongs to the arrow, not to the scan line. Sometimes lettering inside the graph also may act as candidate points at the respective columns. Due to this type of confusion, the points belonging to the columns with a correct spectrum line number, but isolated, are discarded. The remaining classified points are used to form a training set.

After identifying the number of spectra per scanned figure and deciding about the points that belong to these spectra, the remaining work to be done is to connect the segments and get the full spectral plots. For this purpose, the classical "closest distance" approach is used with some modification:

The first modification is to handle the gaps between line segments assigned previously. They are connected piecewise going from left to right and also from right to left. The reason for this decision was this: if a wrong classification is made, filling from either side will limit the wrong classification to half of the gap.

The second modification made was this: if there is no candidate point to classify (for example in the case of broken lines), the last correctly identified y value was assumed for the missing column. To close the gaps, a portion of the image is read from the screen. This portion is 5 pixels wide and 60 pixels above and below the last correctly classified point. The Euclidean distance from these candidate points (i.e., sqrt(x_i^2 + y_i^2)) is calculated, the point with the smallest distance is chosen as the correct point, and in the next iteration this newly classified point is taken as the center of the new window for the array of the candidate points.

As mentioned above, if no point is found at the 5 × 120 pixel window, the previous point is taken to represent the missing portion of the plot.

2.3. Base Line Correction and Identification of the Absorption Bands. The following section refers to transmittance data, so the maximum points mean the base line points and minimum points mean points which belong to the absorption bands. If the IR spectra of polycrystalline or amorphous solids are examined, a property easily recognized is the nonstandard appearance of the spectral plots.[3,4] This effect is mainly due to optical properties of the films. Among the causes are the following:

(a) Sometimes interference patterns are superposed in the spectra.

(b) Some spectra have inclined base lines, and the slope of the base line can be positive or negative.

(c) Scattering losses may cause nonabsorbing regions to have lower transmittance values.

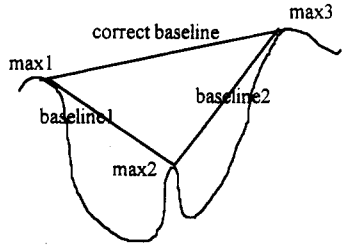

Figure 2. Example of two neighboring bands giving a local maximum.

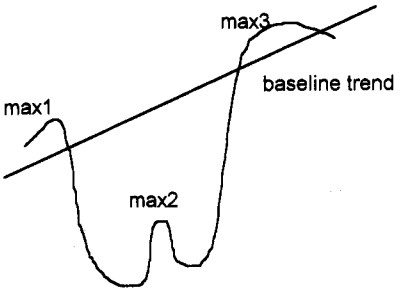

Figure 3. Elimination of the local maximum using the base line trend.

Due to the reasons listed above, the spectra of the solid films always need base line correction.[3] For this purpose, it is accepted that, for the intervals of <1000 cm^{-1}, the base line can be approximated by a straight line. In the base line determination process, first the duplicates of the scan lines were smoothed by applying the moving averages method. Any point y_i is calculated as

$$y_i = (y_{i-1} + y_i + y_{i+1})/3$$

This process is repeated ten times. The reason for the smoothing process is to eliminate the effect of noise while determining the local maximum and minimum points. After the local maxima and minima are found, two vectors which have these points as elements are formed. These points generally belong to the base line, but if some portion at the end points is missing, then the maximum finding logic will not accept these points as the maximum. So the beginning and ending points of the scan lines are also accepted as maximum points. After finding the maxima, a validation process is used. The reason for validating maxima is if a double peak or multiple peaks are present in the spectrum, then the local maxima can cause erroneous results (Figure 2).

As seen in the figure, without a maxima validation process, baseline1 and baseline2 will be drawn, which are not correct base lines. Instead, a base line trend line is used as a demarcation level, and local maxima are discarded from the base line point set as given below.

The base line trend is obtained from the average of the maxima. In the validation process the points like max2 in the Figure 3 are discarded from the maxima list since they are below the base line trend-line.

Finally corresponding points at the original unsmoothed spectrum of the maxima set at hand are connected to form the base line segments. After this drawing, some points are left above the base line. The data are read back from the screen. Then, these points are taken as new maxima (as starting points of the neighboring absorption band), and the base line is redrawn. This process is repeated until no points are left above the base line.

For the spectrum digitization every column is read from the screen; the y value is taken as the difference between the

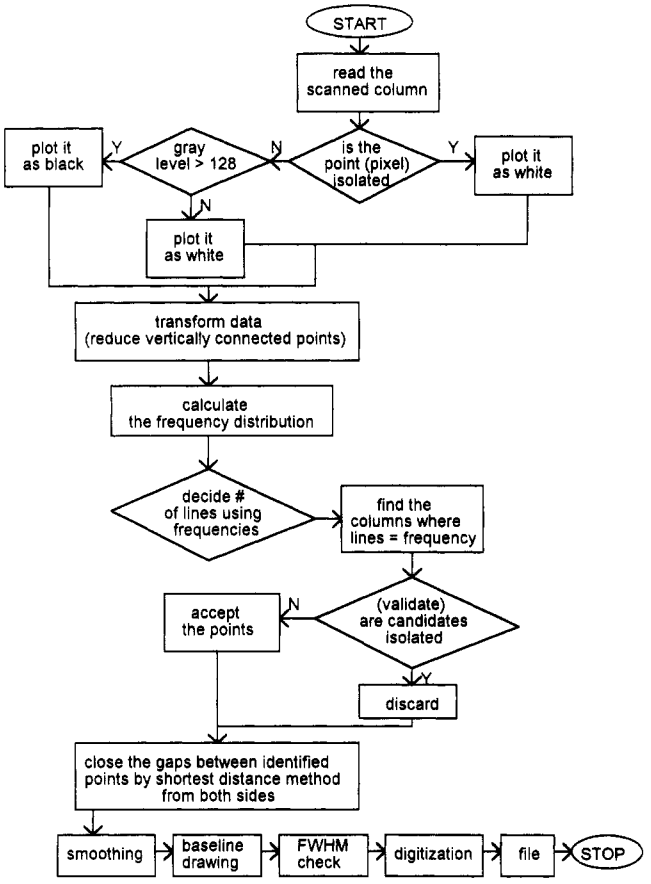

Figure 4. Simplified flow chart of spectrum digitization and conditioning.

y value and the base line. If the point coincides with the base line, that point's y value is accepted as 100.

Before recording the digitized spectrum (Figure 5), a fwhm (full width at half-maximum) check is made to reject the bandlike structures in the spectrum. For each minimum, fwhm values are calculated; then this value, divided by the minimum value, is taken as the demarcation value of the band rejection. For the tested graphs this value is found to be 10, by trial and error method. If the fwhm/peak value > 10, then this structure probably is not an absorption band since an absorption bandwidth must be sharper than this value.

Normalization is made by dividing every point with the minimum of the spectrum and then multiplying by 100. The digitized spectrum is recorded as an ASCII metafile. Each line contains the wavenumber comma transmittance value for that wavenumber, for any exporting to other programs.

Drawing and reading using the same routines gave a byproduct benefit; all the processed spectra have identical wavenumber separations of 5 cm^{-1} regardless of the scanned image size and resolution.

3. RESULTS

3.1. Deciding Number of Scan Lines in a Multiple Scanned Graph. As described in section 2, the number of scan lines is determined from the frequency distribution of points (counting adjacent points as one) in column vectors. Our scanned image set contains 205 graphs (Table 1). The identification in Table 1 is made by excluding the axes. If the regions containing axes are included, the number of scan lines obtained from the computer program is the real number of scan lines plus two.

Table 1. Scan Line Distribution of the Data Set Used for Section 2
Total Evaluated Spectra: 205

no. of scan lines	evaluated spectra	correct scan line guess (%)	no. of scan lines	evaluated spectra	correct scan line guess (%)
1	95	100	6	12	100
2	25	100	7	6	100
3	25	100	8	5	100
4	21	100	9	1	100
5	15	100			

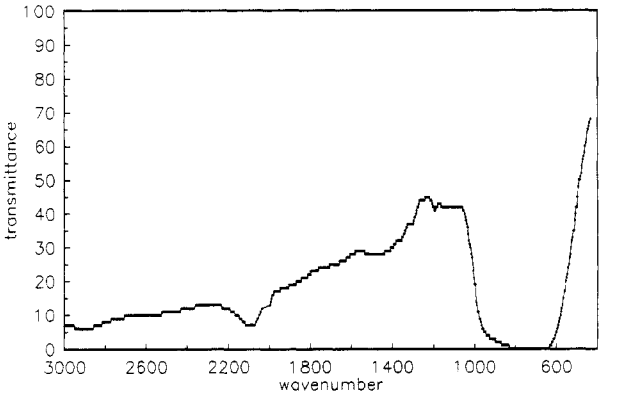

Figure 5. Typical spectrum showing nondefinite band edges.

3.2. Obtaining the Test (Training Set) Group for Classification. Since the number of the scan line determination section worked perfectly, the known set for classifying is obtained from these results. This assumption worked perfectly. By visual inspection it is observed that, before validating the training set, about 3 out of 100 samples contain wrongly classified points. With the validation process, these points are discarded. The remaining parts of the scan lines are found by classifying (using the classifying criteria explained at subsection 2.2) the unknown points between the gaps. This classification process gave correct results except for two spectra. The first of these spectra contains a lot of gaps between the lines. Some of them are large enough that points from the second scan line are taken as the first line; this particular graph also contains vertical lines. The second misclassified image has a very sharp IR spectrum. This group was classified correctly if wider candidate windows were taken (vertically 200 pixels). But, this change degrades identification of the multiple-scan-lined images.

3.3. Base Line Correction. This subsection of the program has drawn the required base line correctly in about 90% of the cases. There are some problem areas. If the spectrum contains uncompleted absorption bands (the remaining part is not given or it is out of the 400–4000-cm^{-1} region), the base line segment drawn for this band may be not correct.

If a band of beginning and ending regions has slopes very near the base line slope, then deciding where the band ends becomes tricky, so the base line segment drawn for the neighborhood of this band may be incorrect. In the base line finding process, if the base line slope is high, then, after the base line drawing, some points are left above the base line, as in Figure 3. The base line is drawn above these points as described in section 2.3. But, this process shifted the absorption maximum 5–50 cm^{-1}, depending on the base line steepness.

3.4. Recording the Absorbance Bands. The bandlike structures are eliminated with use of fwhm checking and also with the integral checking. With the integral checking it is meant that candidate bands with small area integrals are discarded (see Figure 5, similar to ref 5), since most of the small area bandlike structures arise from the noise. Fwhm

checking is described in section 2.3 and is found to be suitable for this group of spectra. After this check only about 5%, of the bandlike structures are left. Integral checking eliminated the noise along with the very small bands (if it exists).

4. CONCLUSION

A set of simple algorithms are proposed and tested to digitize the IR spectral data for polycrystalline and amorphous silicon alloys. This method and the procedures were required for a task of library formation for the silicon alloys. But, this method is also suitable for other x–t type single valued spectrograms of any measurement. The only change to be made is entering new acceptable fwhm values. Deciding the scan line number by using the frequency distribution worked perfectly.

Obtaining the training set for the classifying process by using the frequency distribution of the columns seems to be suitable for this type of digitization work. But, the validating process may not work on the more complex, faint, and broken-lined graphs. Even though it is not observed in the tested 218 examples, some graphs may contain columns with three or more adjacent points having frequency distributions equal to the number of scan lines in the graph. The proposed system needs more testing to be done for checking the suitability of the validation criteria.

Base-line correction subunits based on the mentioned algorithms are the most controversial part of this work. As given in the results section, if an incomplete band is present in the graph, digitizing using base line data causes the band to shift to higher values if the base line slope is negative or to shift to lower values if the slope is positive. The shifting value depends on the sharpness of the band: the higher the sharpness, the higher the shifting. Since the signal conditioning is a must for this type of inclined graph, a correction method for this effect is needed.

The proposed integral checking method for eliminating the bandlike structures arising from noise, interference, and optical inhomogeneities of the films worked well in the most cases. But, with this process some of the small real bands are also discarded. An algorithm to separate small bands from the noise would be helpful.

ACKNOWLEDGMENT

This work is supported by TUBITAK (Turkish Scientific and Technical Research Council) with Grant EEEAG-78.

REFERENCES AND NOTES

(1) Kavak, H.; Esen, R. Spectrum Comparison of IR Data Taken from Difference Spectrometers with Various Precision. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 595–7.
(2) Jurs, P. C.; Isenhour, T. L. *Chemical Application of Pattern Recognition*; Wiley-Interscience: New York, 1975; pp 9–16.
(3) Tomellini, S. A.; Saperstein, D. D.; Stevenson, J. M.; Smith, G. M.; Woodruff, H. B.; Seeling, P. F. Automated Interpretation of Infrared Spectra with an Instrument Based Minicomputer. *Anal. Chem.* **1981**, *53*, 2367–9.
(4) Ruprecht, M.; Clerc, J. T. Performance Analysis of a Simple Infrared Library Search System. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 241–4.
(5) Beyer, W. *Appl. Phys. Lett.* **1989**, *54*, 1668.