Investigation on Automatic Data Acquisition of Well Logging Curve based on Color Recognition

Han Wang¹, Zhihui Ye¹*, Dong Chen², Dandan Zhu³, Shouding Lí⁴,⁵,⁶

¹College of Safety and Ocean Engineering, China University of Petroleum-Beijing, Beijing, 102249, China
²State Key Laboratory of Petroleum Resources and Prospecting, College of Petroleum Engineering, China University of Petroleum-Beijing, Beijing, 102249, China
³Beijing Key Lab of Petroleum Data Mining, Department of Computer Science and Technology, Beijing, 102249, China
⁴Institute of Geology and Geophysics, Chinese Academy of Sciences, Key Laboratory of Shale Gas and Geoenineering, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, 100029, China
⁵University of Chinese Academy of Sciences, Beijing, 100083, China
⁶Institutions of Earth Science, Chinese Academy of Sciences, Beijing, 100029, China

*Corresponding author’s e-mail: zhihui.ye@cup.edu.cn

Abstract. Well logging data are important to evaluate rock and fluid properties underground. Many well logging data, as functions of well depth, are conserved as figures in the literature. Digitizing techincs should be applied when utilizing the well logging data. In order to address this problem, a method was proposed to automatically identify and acquire the well logging data from well logging curves. Based on the color recognition, this algorithm is able to extract curve features based on pixels. Comparing to other data digitizing methods, the background grid lines could be easily filtered and removed. The algorithm is verified by digitizing some typical well logging curves. This main contribution of this work is to provide a new and efficient way for data digitizing from well logging curves.

1. Introduction

Well log curves are important parameters to describe the properties of rock and fluid encountered during drilling. It has a very important reference for the selection of drilling parameters and the formulation of oilfield development plans. Well logging data are widely used for reservoir evaluation in petroleum industry. Most logging data exist in the figure format in the literature. A reliable digitizing technic is required to utilize the logging data.

In order to conduct data analysis, researchers need to extract the data of these curves, then store and use these data[1]. But now commonly used data extraction software, e.g. Engauge, g3data, GetData, Dagra, DataThief, dcsDigitiser, DigitizeIt do not work well for well log curve, especially with gridlines as background. The advantages and disadvantages of these products are described below:
Dagra mainly identifies the curves by curve trend. First, extract the frame from the figure. Then select and name the horizontal and vertical coordinates and the coordinate range is determined. Finally, the data is extracted by selecting the key points on the curve and pulling out the influence lines on both sides of it. The method of drawing influence line to determine the general trend of data curve is very slow, but it can improve the accuracy of data collection.

DataThief, GetData, g3data, Graph Click etc. generally identify data curves by selecting some feature points to get the regional recognition. Other steps are similar to Dagra, and there are many softwares of this mode. One disadvantage of these software is that if there are a large number of grid lines or interference lines in the figure, it will greatly affect the accuracy of the extracted data and even make the data difficult to be extracted.

Compared to the above software, the DigitizeIt and im2graph only have some additional functions. The DigitizeIt can automatically pick up data in scatter plots. The im2graph has the function of recognizing the text in the figure and making the table head and the name of x-axis and y-axis. These new small functions can make data acquisition of well logging curve higher efficiency, but still can’t effectively remove interference of grid lines to get accurate results.

In general, the widely used data fetching software has the following small problems:
1. The obtained data may have deviation compared with the data curve.
2. Needed clicks the data area repeatly to select the data point, which is time consuming.
3. The anti-interference ability to the auxiliary grid is poor.[2]

Considering the above problems, we have designed this program, which is specifically targeted to the data image with discrepancy on color region.[3] It is more convenient, fast and accurate to take out the data of all data points in the data graph and normalize the data to the same group. Compared with the commonly used software, this method has the characteristics of higher precision, time-saving and labor-saving, but it’s not good for the curves without color difference.

2. Methods

2.1. Processing process

In this paper, logging curve data recognition is realized, and the processing flow of it is shown in figure 1.

2.2. Image preprocessing

First, adjust the photo or scan image from unstandard rectangle to a standard rectangle. Next, click and select the part of curve that is boxed in the required coordinate. The system will crop out these pixels and store them in a separate matrix. Then, remove some of the noise in the matrix [4]. Finally, save the maximum and minimum values of the data (the X-axis and Y-axis’s coordinate range) [5].
2.3. Feature pixel recognition
Store the select data matrix as an RGB image and open the image in the image interface. Next, the operator needs to select a rectangular area in the image interaction interface with the features of data pixels to be extracted, click the points in the upper left and lower right corner of the rectangular area. Store the color eigenvalue of this area in another color matrix. The code will automatically remove all the white pixels of this matrix and count the color characteristics of the remaining pixels, which will be take as the initial eigenvalue of the data and assigned it to three variables r, g and b.

If the background pixel is not white, it is necessary to select a part of area of the background color pixel. When extracting the color features of the curve, it is necessary to delete these pixels, and then the remaining pixels are taken as the average of the color features.

Next, judge the r, g and b respectively, and get a value range based on the curve's color feature after a series of processing of the judged color feature value. The color feature value obtained here will be used as the basis for subsequent data collection.

The range of the color features is determined by the three color channels R, G and B, and the value range of these color channels are from 0 to 255. Through reasonable division, the main body of the three-color channel can be divided into the following areas to distinguish the color differences: include black, red, green, blue, yellow, cyan and magenta. These color ranges are somewhat overlapping, because many of the curves's color seen by the naked eye are not the same with accurately identified by the computer. The initial value of the color feature range will be selected based on the previously extracted pixel color feature, and the selected color range is also based on this value. Most curves can extract colors in this range. In case of unrecognized color region, the value range of the three-color channel can be manually adjusted until the complete curve is taken out. After a large number of curve extraction attempts, it is proved that the division of the color channels is reasonable. The principle of feature pixel recognition is shown in figure 2 below:

![Feature pixel recognition](image)

**Figure 2. Feature pixel recognition.**

2.4. Remove grid interference
The interference of dark or light grid lines often exists in the extraction of drilling curves, which must be remove to obtain the accurate data [6]. The first should be removed is the interference of light grid lines [7]. By setting the color eigenvalues, all the pixels which color eigenvalues are larger than the color range are filtered out to filter the light grid lines [8]. Then the dark grid lines should be filtered. First, the data on the dark grid line and data curve were taken out together, and then the entire row element with the proportion of pixels greater than n (n adjustable) on each row and column of the
assignment matrix was removed as the background color. The result of removed grid lines interference is shown in figure 3 below:

![Figure 3. Remove grid lines interference.](image)

2.5. Identify all data points on the graph
First, the characteristic values of the extracted data curve color are used to judge which color region it belongs to. Secondly, after determining the color region, select the feature range of the color region, and list all the pixel’s coordinates in a list through this feature range. By adjusting a variable, adjust the color range of the degree of relaxation. The less loose color range is, the fewer pixels will be select and vice versa. Finally, turn the selected point set into a white area on the grayscale image according to weather corresponding position and observe whether this line is the desired data part. If not, the degree of accommodation can be adjusted until appropriate [9]. The data comparison figure [10] is shown in figure 4 below:

![Figure 4. Grid data extraction graph.](image)
2.6. Data vacancy filling
Some of the acquired pixels are discarded and some new pixels are added to complete the data graph. First, identify the pixel coordinates of the extracted eigenvalues on each row and calculate the average value to find the most appropriate location of the points in this coordinate. Secondly, identify the pixels at the end and the begin of the curve to see if there is a position with null values. If there are some null values, the coordinate range will be set to the position of the data point nearest to the null value. Finally, if the null part (i.e. the broken part of the dotted line) appears in the middle part of the curve, the null part will be completed according to the data coordinates at both ends of the null value. The completion method is mainly based on the interpolation method, which analyzes the area curves before and after the broken part and then interpolates the null part [11]. The data obtained after interpolation will be continuous and complete.

2.7. Comparison and storage
Draw the list of coordinates obtained from the previous step as a broken line and compared with the original curve. After the comparison, the point coordinates were transformed into data values at the same proportion, and the data values were stored in CSV in the form of columns. The comparison flow chart of extracted data is shown in figure 5 below:

![Figure 5. Data extraction flow chart.](image)

2.8. Uniform interval of curve data
After several logging curves of the same well section are counted by this method, data at the same Y-axis (for example, at the same depth) need to be processed at equal intervals and their intervals can be set before input data analysis tools. After the data of several curves are input, the common depth data part is selected and the part on each curve is taken out. The extracted data are then interpolated based...
on a certain interval, so that groups of data at a same interval can be get. This same interval data can be directly used to do data analysis and data model construction.

3. Conclusion
In order to improve the accuracy and efficiency of the data extraction from the curves, this method based on color recognition was invented. Through the analysis of the data acquisition effect of this method and the comparison of current commonly used software, the following conclusions can be obtained:

- The accuracy of data extracted by this method is higher than extracted by conventional methods. Because it is based on color recognition, it can get high efficiency in the data graph with chromatic aberration. As long as the value range of the horizontal and vertical coordinates is selected and the region color is selected through two points, all the data on the whole curve can be extracted.

- This model does not work well on multi-curve graphs with very close color difference, because it is based on color recognition. In this case, data of multiple curves will be identified together, and inaccurate data points will be obtained after processing.

- In the case of most data graphs, this method has higher precision and higher efficiency than the conventional software, and has a strong advantage in the processing of grids. This approach has great potential.

Acknowledgements
This study was funded by the strategic science and technology initiative of the Chinese academy of sciences (category A). The project numbers are XDA14040402, XDA14040401 and XDA14040404.

References
[1] Yan Guangwei, Hu Haitao, (2017) Identification and data extraction of time series curves in scanned images. Computer Research and Development, 44(Suppl.):251-256.
[2] Qian Peng, Liu Ming, Li Yanqi, et al.(2015) Development and application of logging curve identification and digital system[J]. Oil and Gas Reservoir Evaluation and Development, (2): 67-72.
[3] Jan Erik Solem, (2014) Computer Visual Programming, O'REILLY, People's Posts and Telecommunications Press, Beijing.
[4] Luo Feng, Yu Yanmei, Wu Xiaohong, et al. (2005) Background denoising method in curve vectorization[J]. Journal of Computer Applications, (4): 929-931.
[5] Luo Feng, Yu Yanmei, Liu Fei, et al. (2005) Study on Pretreatment Methods in Logging Curve Vectorization[J]. China Journal of Test and Measurement, (3): 40-41, 64.
[6] Liu Wei. (2009) Research on vectorization technology of logging drawings [D]. Harbin Engineering University, 1-65.
[7] Li Zengbing, Zhang Xiaofan, Tang Lanlan, et al. (2006) Research on automatic vectorization algorithm for linear volume images[J]. Modern Surveying and Mapping, (3): 11-12,16.
[8] Yu Xuejun, Peng Lizhong. (2002) A New Algorithm for Extracting Contours of Binary Image Curves[J], Chinese Journal of Image and Graphics, Series A, (3): 272-275.
[9] Zuo Hang. (2003) Automatic identification and extraction of logging curves [D]. Sichuan University, P63.
[10] Bijan Mostaghel, (1999) Processing and Interpretation of geophysical well logs from well KJ-32, krafla geothermal field, Ne-Iceland, The United Nations University’s Reports, Number 8
[11] Yu Yanmei, Luo Daisheng, He Xiaohai, et al. (2008) Design and development of intelligent digital system for logging curve[J]. Computer Engineering and Design, (24): 6405-6406,6409.