# Processing Algorithm of Irregular Table Image in Tobacco Package Based on Dual-coding Difference of Gaussians Method

**LyuZhigang, AssociateProfessor**
**Wang Hongxi, Professor**
**Li Liangliang, Ph.DCandidate**
**Wang Peng, Professor**
**Li Xiaoyan, AssociateProfessor**

*LyuZhigang, Associate Professor in Control Science and Engineering, School of Mechatronic Engineering, Xi'an Technological University, Xi'an 710021, China; School of Electronics and Information Engineering, Xi'an Technological University, Xi'an 710021, China. Wang Hongxi, Professor inMechatronicandElectronicEngineering, School of Mechatronic Engineering, Xi'an Technological University, Xi'an 710021, China. Li Liangliang, Ph.D Candidate inMechatronicandElectronicEngineering, School of Mechatronic Engineering, Xi'an Technological University, Xi'an 710021, China. Wang Peng, Professor inControlScienceandEngineering, School of Electronics and Information Engineering, Xi'an Technological University, Xi'an 710021, China. Li Xiaoyan, Associate professor inCommunicationEngineering, School of Electronics and Information Engineering, Xi'an Technological University, Xi'an 710021, China.Correspondence author: Wang Hongxi;375261456@qq.com*

**Objectives: Currently, in a large number of print-out report documents from tobacco package, there exist irregular phenomena such as discontinuous vertical lines, misplaced frame lines and multi-page tables. Thus, the existing table recognition algorithm cannot be adopted to perform digital identification. In order to solve this problem, this paper proposes a table image processing algorithm based on the dual-coding difference of Gaussians iterative clustering. Firstly, the method of local regional sub-block is used to the skew correction threshold to conduct image correction. Secondly, the corrected images are coded by rows and columns, and 2D image features are transformed into 1D image features. Thirdly, the Gaussian differenced operation is adopted to obtain effective characteristic matrices that are stable and easily distinguishable. Then the iterative clustering analysis is performed to obtain the feature values of effective frame lines. Fourthly, after finishing the tasks, such as the table positioning, inner structure reconstruction, and text information identification, the dichotomy judgmentsof the integrity of multi-page tablesare realized according to the local pixel features. Finally, the text information inside the local regions and the reconstructed regions are merged, and the digital reproduction of the multi-page tables is realized. To validate the effectiveness of the proposed algorithm, an experiment in the sample set containing 12,840 table images with different resolutionsis carried out. The average detection accuracies of table positioning, table cell reconstructionand multi-page incompleteness are 98.95%, 99.80%, and 95.85%, respectively. The experimental results show that the proposed algorithm is simple and effective, and can accomplish the digital reproduction of irregular tables.**

**Keywords:Image dual-coding; Difference of Gaussians; Irregular table; Tobacco package**
*Tob Regul Sci.™ 2021;7(5):1170-1188*
*DOI: doi.org/10.18001/TRS.7.5.35*

## INTRODUCTION

It is of significance for the management of tobacco package to perform digital and text information extraction of the print-out documents. Due to the limitation of early tabulation techniques, the early print-out documents contain irregular tables, which are manifested as the discontinuous vertical lines, misplaced frame lines and multi-page tables. At present, some OCR softwarein the market can identify spreadsheet, but excessively rely on the API or online SDK provided by the cloud platform. Thus, it is not suitable for them to process the documents with a certain security classification. Moreover, these OCR software can only analyze the current page's complete and regular table. For the irregular table on different pages, they do not help at all. Therefore, studying the offline processing algorithm of irregular tables has become the key to solve the above problems. In recent years, the processing algorithms of regular tableshave significantly and rapidly improved. The contrastive analyses of advantages and disadvantages for various algorithms are as below:

Kuang Zhen et al. adopted the projection method to extract the horizontal and vertical line coordinates, and constructed table feature points, which could realize the recognition of vote tables [1]. Duan Lu et al. proposed a fast correction method and adopted the projection method to realize the line segmentation of questionnaire tables[2]. Bai Wei et al. used a clustering method based on run length and collinear lines and improved the recognition rate of complex tables. However, the time consumption of the algorithm was increased [3]. Anukriti Bansal proposed a table extraction algorithm based on the fixed point model[4]. However, the algorithm has high complexity and is not easy in practical applications. Manabu Ohta proposed a table cell detection method based on the table model, which could realize the table image detection of handwritten documents and the recognition of table cell text content. However, the table-driven model is rather complicated, and the practical feasibility and robustness are significantly limited [5].Dong Xiao proposed a table recognition method based on boundary detection. The processing objects are regular tables [6]. Qiao Kang Liang adopted the improved Sauvola robust binarization algorithm, which could better process the images with uneven lighting and blurred imaging, and extracted the table horizontal and vertical lines based on the morphological detection operators. However, its robustness and effectiveness have not been verified yet, and the morphological detection cannot realize the self-adaption of the threshold [7]. Jihu Kim used the method of feature extraction and grid cognition to detect the table. The method has a fast processing speed and is superior to the methods that usually use classifiers and deep neural networks to extract the features. However, it can only process the regular tables [8]. Smite Pallavi proposed an algorithm for detecting and extracting multiple tables from optical character recognition (OCR) documents. The algorithm adopted the method of traditional construction level and vertical structure elements to realize table reconstruction. The experiments showed that the method has some defects[9].

Muhammad Ali Shahzad proposed an alternative method of solving the complex deep learning problem. The method has certain effectiveness for coding the traditionally hand-made block structure features, text and digital features as well as the image's background features [10]. Yibo Li and Pau Riba proposed a GAN-based feature generator that detected the regular tables and had a certain practical effect [11,12]. Shubham Singh Paliwal proposed a deep learning model TableNetfor extracting the table data from scanned document images. The method simultaneously performs table detection and structure recognition. Shubham Singh tested the method on ICDAR 2013 dataset and showed excellent test results [13].Yilun Huang proposed an improved YOLOV3 algorithm used for table detection. The experimental results showed certain practical feasibility [14]. Isaak Kavasidis proposed a saliency-based convolutional neural network to conduct multiscale reasoning for visual clues and used the conditional random field (CRF) to position the table and chart in

digital documents. The method was tested on ICDAR 2013 dataset and showed excellent test results [15]. Devashish Prasad proposed an improved end-to-end method based on deep learning. The method used a single convolutional neural network model to solve the table detection and structure cognition problems. However, for the reconstruction of the table internal structure, it still adopted the traditional method of extracting horizontal/vertical lines and solving mask intersection. For the problem of being hard to determine the high-resolution image threshold, the method proposed by Devashish Prasad has certain limitations[16].

Literature [1-9] adopted the traditional method to realize table recognition, which has the problems such as weak self-adaptability, and poor robustness. Based on the table recognition of the deep learning model, Literature [10-16] finally obtained the table boundary coordinates. The algorithms rely on the hardware device performance and pre-training test of a large number of datasets, which significantly limited the application universality. Compared with the traditional table layout analysis method, the deep recognition method has improved the test accuracy, but for the reconstruction of table structure, it still adopts the traditional morphological detection algorithm. Meanwhile, all the existing methods focus on the studies of regular tables, but for the irregular tables with the phenomena such as discontinuous vertical lines, misplaced frame lines and multi-page tables, fewer studies have been conducted.

In order to resolve the above-mentioned problems, this paper proposes an image recognition algorithm based on solving the table and character elements recognition problems in tabular files. At the same time, the paper puts forward the judgment and joint algorithm of the multi-page table and realizes the context relevance of multi-pages and batch processing, therebyguaranteeing the completeness of data structure and improving the work efficiency.

## RELATED WORK

The irregular tables studied in this paper refer to the tables with irregular phenomena such as discontinuous vertical frame lines, misplaced vertical frame lines and multi-page tables. The samples are shown in Figure 1:

| Name | Theoretical value | Actual value | Error | Conclusion |
|---|---|---|---|---|
| | | | | Qualified |
| | | | | Qualified |
| | | | | Qualified |

a. Discontinuous vertical frame lines

| Name | Theoretical value | Actual value | Error | Conclusion |
|---|---|---|---|---|
| | | | | Qualified |
| | | | | Qualified |
| | | | | Qualified |
| | | | | Qualified |

b. Misplaced vertical frame lines

| | | | | |
|---|---|---|---|---|
| Test | | 0.000 | 0.000 | |
| t1(s) | | 4.568 | 0.000 | |

| | | | | |
|---|---|---|---|---|
| t2(s) | (t2=0) | 0.000 | 0.000 | |
| t3(s) | (t3=0) | 0.000 | 0.000 | |

c. Multi-Cont-Page

**Figure 1 Samples of irregular tables**.

The vertical frame lines in Figure 1 (a) are discontinuous dotted lines, and their vertical frame lines are thin and short. Therefore, it is not easy to directly perform vertical frame line detection. The 2-6th vertical frame lines in Figure 1 (b) are misplacedto different extents and are also difficult to be detected. The frame lines in Figure 1 (c) are incomplete, but the sample has content relevance between the up and down tables(i.e., the multi-page table). It should be noted that the first and the last tables on the current page may have the phenomenon of "the multi-page table". The studies on irregular table characteristic analysis in this paper are shown in Table 1.

**Table 1**
**Irregular Table Characteristic Analysis**

| Resolution | Vertical framelinecharacteristics | Multi-pagecharacteristics |
|---|---|---|
| 75-300dpi | Vertical frame line is a discontinuous dotted line | Top or bottom multi-page |
| 75-300dpi | Vertical frame line is discontinuously misplaced | Top or bottom multi-page |

For the irregular tables above, the disadvantage of the existing algorithms is that they cannot correctly finish the structure reconstruction. Given the table's inner structure reconstruction, the existing algorithms perform relevant research all based on table frame line or horizontal/vertical lines intersections. This leads to the phenomena such as undetectable intersections and location misjudgment of intersections, during the detection of intersections, as shown in Figure 2.

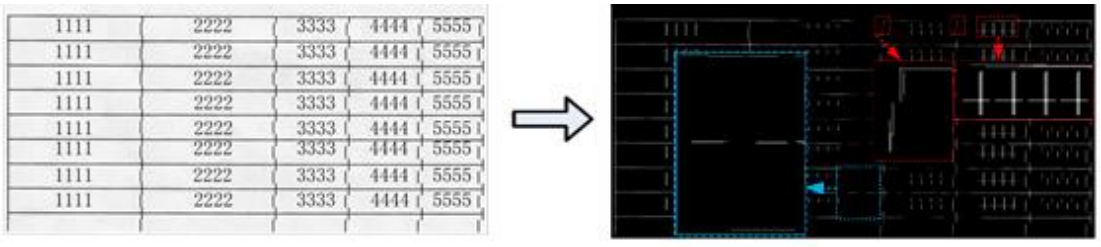



**Figure 2 Output Images of Intersections Detection Errors by the Traditional Algorithms.**

In the traditional algorithms, the intersection superposition fusion image matrix $F_{H+V}$ of horizontal/vertical lines is defined as:

$$F_{H+V} = H_{lines}/2 + V_{lines}/2 \quad (1)$$

where, $H_{lines}$ is the horizontal line detection output image matrix, and $V_{lines}$ is the vertical line detection output image matrix. In Figure 2,

the intersection $F_{H+V}$ is directly displayed in the form of bright spots.

Since the irregular table cannot be well recognized by the existing methods, this paper proposes a processing algorithm based on the dual-coding difference of Gaussian iterative clustering. In main work in this paper is summarized as follows:

(1) An irregular frame line processing algorithm based on dual-coding differenced iterative clustering is proposed to process the text reports with multiresolution and various imaging qualities;

(2) A completeness processing algorithm of the table across pages based on local pixel feature classification is proposed;

(3) Atext information recognition model is constructed,realizing the digital reproduction of report text information.

## ESTABLISHMENT OF THE ALGORITHM MODEL

After the original images are input to process graying, skew correction is performed for the gray images. Then, the proposed model is employed, mainly including vertical/horizontal image coding (H/V IC), difference of Gaussian (DOG), iterative difference clustering (IDC), table structure reconstruction (TSR), and text info recognition (TIR). The H/V IC mainly aims to code the 2D image features after skew correction into 1D image features. After the DOG is used for the 1D image coding feature, the local features of the table frame lines become prominent. The IDC mainly performs the clustering process for the data after DOG processing to realize the positioning of the table horizontal/vertical frame lines. The TSR finishes the table reconstruction and multi-page judgment by analyzing the table horizontal/vertical frame lines. The TIR adopts a double projection statistical method to position the local text information and uses the convolutional cyclic neural network model to perform text information recognition, which has provided the foundations for realizing the digital reproduction of tables. The flowchart of the algorithm proposed in this paper is shown in Figure 3.

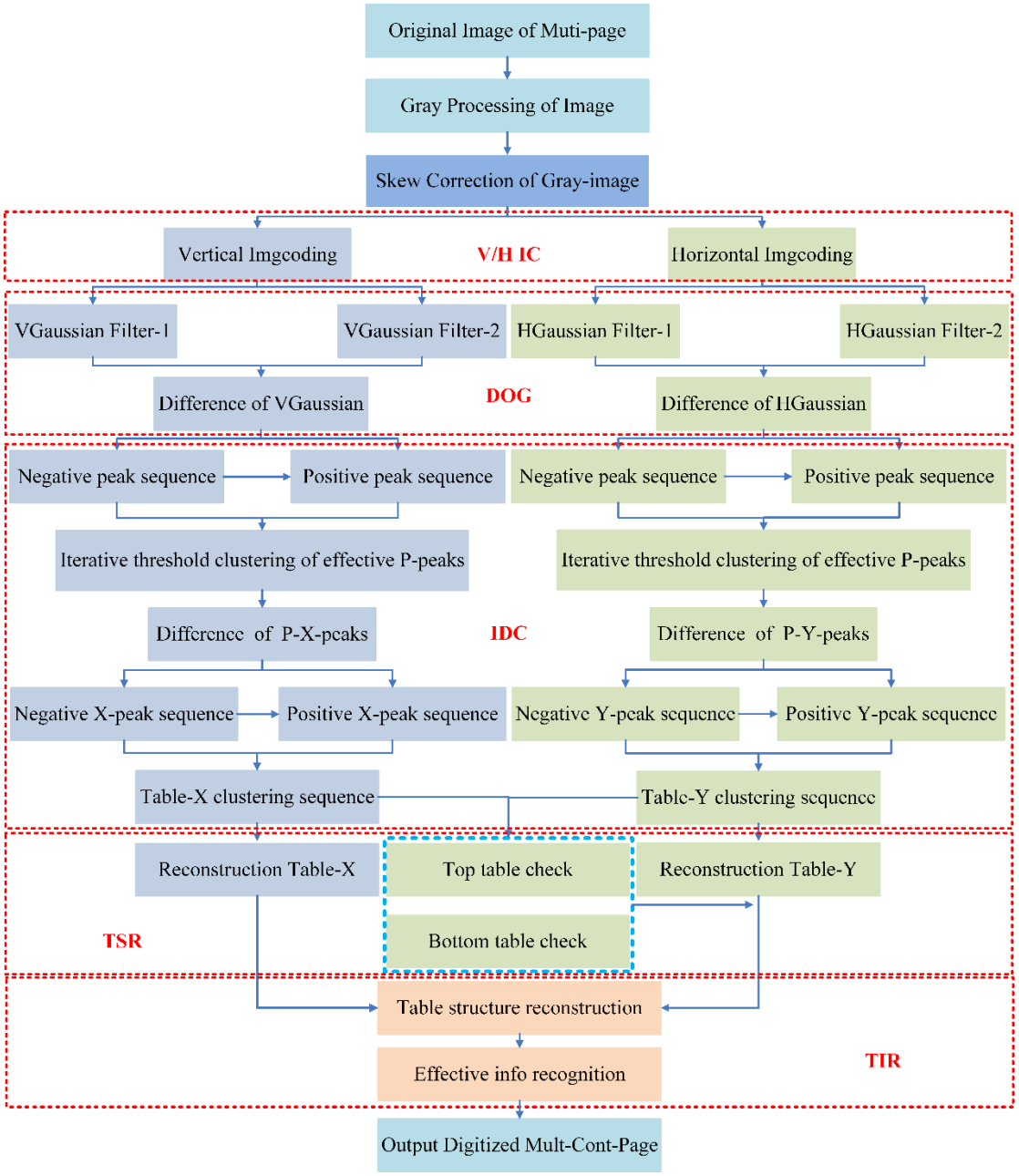



**Figure 3 Flowchart of the Algorithm Proposed in this Paper**

**Skew correction algorithm based on local features**

In this paper, a global image correction method based on the local region features is designed. Through skew correction, the table to be processed with informatization is corrected. Firstly, the multi-resolution images are grayed, and then are processed with the OSTU binarization. By making the regional sub-block rules, the binary images are extracted by sub-blocks to obtain the specified central region images. In the specified central region images, the morphological method is used to extract the horizontal lines and then Hough reconstruction is performed to improve the accuracy of horizontal frame lines. Then, the slope average of the local horizontal lines is solved, and the skew correction of the gray image is accomplished. At the same time, the mean value of therow spacingata specified central region in the image is solved, which can provide the basis for the consequent TSR. The main steps are shown in Table 2.

**Table 2**
**Skew Correction Algorithm**

Input: original image; Output: the corrected image and row spacing average

1: Input image graying, and perform OSTU binary processing

2: Make regional sub-blocks according to the image width and height, and extract the central region

3: Construct the structural element $S = size(width/10,1)$, and pre-extract the horizontal lines for the central region

4: Perform Hough reconstruction because of the horizontal lines

5: Solve the slope average of horizontal lines, and calculate the local average rotation angle

6: Calculate the average of 2D rotation affine transformation matrix with row spacing, and complete the skew correction of gray images

In the skew correction algorithm above, the key supplements of steps are specified as follows:

(1) Extraction rule of regional sub-blocks

The height and the width of the input image are set as $height$, and $width$, respectively. The diagram of regional sub-blocks is shown in Figure 4.

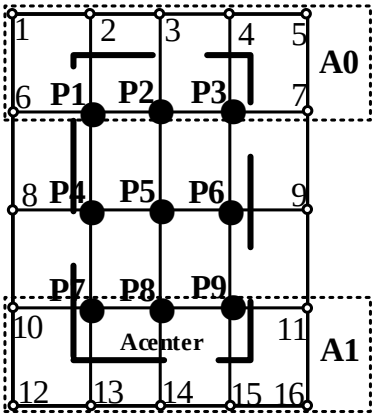



**Figure 4 Diagram of image sub-blocks.**

The coordinates of P1-P9 in Figure 4 are shown in Table 3.

**Table 3**
**Coordinates of Sub-Blocks**

| Point | X | Y |
|---|---|---|
| **P1** | $width/4$ | $height/4$ |
| **P2** | $width/2$ | $height/4$ |

| | | |
|---|---|---|
| **P3** | $width_{/4+}\ width_{/2}$ | $height_{/4}$ |
| **P4** | $width_{/4}$ | $height_{/2}$ |
| **P5** | $width_{/2}$ | $height_{/2}$ |
| **P6** | $width_{/4+}\ width_{/2}$ | $height_{/2}$ |
| **P7** | $width_{/4}$ | $height_{/2+}\ height_{/4}$ |
| **P8** | $width_{/2}$ | $height_{/2+}\ height_{/4}$ |
| **P9** | $width_{/4+}\ width_{/2}$ | $height_{/2+}\ height_{/4}$ |

In Figure 4, P1-P9 are the intersections of the input image lines with 16 equal divisions; Region A0 is the local region consisting of Points 1, 5, 6, and 7; Region A1 is the local region consisting of Points 10, 11, 12, and 16; and Region Acenter is the local region consisting of the midpoints of Lines 2-P1, 4-P3, 13-P7 and 15-P9.

(2) Skew correction

The input image matrix is set as $X$, the image width is set as $width$, the corrosion and expensive matrices are set as $S_1 = size(width/10,1)$ and $S_2 = size(3,3)$, respectively. According to Equation (2), the morphological method is adopted to perform once corrosion and twice expansion operations in the Acenter region, and obtains the image matrix $X_{horizontal}$ with only horizontal lines.

$$X_{horizontal} = ((X \Theta S_1) \oplus S_1) \oplus S_2 \quad (2)$$

Then, Hough transformation is adopted to conduct reconstruction processing for horizontal lines. The average of all lines' slopes in the image matrix $X_{horizontal}$ can be calculated, and also the average rotation angle can be obtained. Then, a 2D rotation affine transformation matrix can be obtained, and thus skew correction of the gray image can be finished.

(3) Calculation of row spacing average

In step (2) above, during the reconstruction of the Hough line for $X_{horizontal}$, it can obtain the horizontal line's vertical coordinates set $Y_{Acenter}$. According to Equation (3), the lines are clustered, and the vertical coordinates set of the lines satisfying clustering conditions is in set $Y_{Acenter\_h}$.

$$Y_{Acenter\_h} = Y_{Acenter}(i+1) - Y_{Acenter}(i) < 20(i >= 0) \quad (3)$$

Then, according to Equation (4), the average coordinates of the same cluster are solved, and the row spacing is calculated. $\{Y_{Acenter\_h}(i+1)/(i+1)\}$ represents the average coordinates of the $i+1$ th type of lines, and $j$ represents the number of lines.

$$MRL = \begin{cases} \dfrac{\sum_{2}^{j}(\{\frac{Y_{Acenter\_h}(i+1)}{i+1}\} - \{\frac{Y_{Acenter\_h}(i)}{i}\})}{j-1} & (i >= 1, j > 1) \\ 100 & (j = 1) \end{cases}$$
$$(\{\frac{Y_{Acenter\_h}(i+2)}{i+2}\} - \{\frac{Y_{Acenter\_h}(i+1)}{i+1}\}) - (\{\frac{Y_{Acenter\_h}(i+1)}{i+1}\} - \{\frac{Y_{Acenter\_h}(i)}{i}\}) \in (0,20) \quad (4)$$

**Dual-coding difference of Gaussian construction method**

Through the aforementioned skew correction processing, the currently processing table image can be regarded to be horizontal and vertical. At the same time, considering the irregular table's features in this paper, the 5x5 input image is taken as an example and the image coding method is shown in Figure 5. In the figure, the numerical values in the coordinate space are the pixel values of various points. The last numerical value of each row/column is the cumulative sum of the current row/column's pixel values.

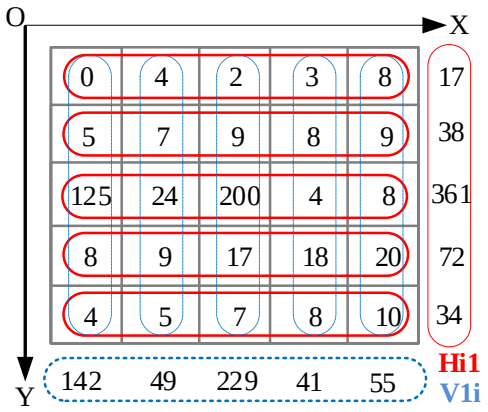



**Figure 5 Coding diagram.**

The input image matrix is set as $I_{M\times N}$, and according to Equations (5) and (6), 2D image features are transformed to be 1D image features, marking as $H_{M\times 1}$ and $V_{1\times N}$, respectively. According to Equations(7) and (8), the updated $H_{M\times 1}$ and $V_{1\times N}$ can be obtained, where $H_{i1}$ and $V_{1i}$ are the $i1$ th and $1i$ th data in $H_{M\times 1}$ and $V_{1\times N}$, respectively; $H_{Mean}$ and $V_{Mean}$ are the mean values of $H_{M\times 1}$ and $V_{1\times N}$, respectively; while $H_{std}$ and $V_{std}$ are the standard deviation of $H_{M\times 1}$ and $V_{1\times N}$, respectively.

$$H_{i1}=\sum_{j}^{N} I_{ij} \quad (5)$$

$$V_{1i}=\sum_{j}^{M} I_{ji} \quad (6)$$

$$H_{M\times 1}=\frac{H_{i1}-H_{Mean}}{H_{std}} \quad (7)$$

$$V_{1\times N}=\frac{V_{1i}-V_{Mean}}{V_{std}} \quad (8)$$

The coded 1D image features $H_{M\times 1}$ and $V_{1\times N}$ are substituted into Equations (9) and (10) to perform horizontal and vertical Gaussian difference processing, respectively, in which the standard deviation is $\delta 2>\delta 1$.

$$H_{DOG}=(G_{\delta 2}(x,y)\times H_{M\times 1}-G_{\delta 1}(x,y)\times H_{M\times 1})\times 200 \quad (9)$$

$$V_{DOG}=(G_{\delta 2}(x,y)\times V_{1\times N}-G_{\delta 1}(x,y)\times V_{1\times N})\times 200 \quad (10)$$

where $G_{\delta x}(x,y)$ is shown in Equation (11):

$$G_{\delta x}(x,y)=\frac{1}{\sqrt{2\pi\delta_x^{\,2}}}\exp\left(-\frac{x^2+y^2}{2\delta_x^{\,2}}\right) \quad (11)$$

The Gaussian differenced pseudocode flow of 2D image features transformed into 1D image features is shown in Table 4.

**Table 4**
**Gaussian Differenced Pseudocode Flow of Image Coding**

| **Input: gray $I_{M\times N}$; Output: Differenced image after horizontal/vertical coding** | |
|---|---|
| **1:** | $H\leftarrow sum_j^N I_{ij}$, $V\leftarrow sum_j^M I_{ji}$ |
| **2:** | $H_{Mean}\leftarrow sumH/N$, $V_{Mean}\leftarrow sumV/M$ |
| **3:** | $H_{std}\leftarrow sqrt(sum[(H-H_{Mean})\wedge 2]/(N-1))$, $V_{std}\leftarrow sqrt(sum[(V-V_{Mean})\wedge 2]/(M-1))$ |
| **4:** | $H\leftarrow H-H_{Mean}$, $V\leftarrow V-V_{Mean}$ |
| **5:** | $H\leftarrow H/H_{std}$, $V\leftarrow V/V_{std}$ |
| **6:** | $H_{g1}\leftarrow Gaussian(H,\delta 1)$, $H_{g2}\leftarrow Gaussian(H,\delta 2)$ |

| | |
|---|---|
| **7:** | $V_{g1} \leftarrow Gaussian(V,\delta 1)$ , $V_{g2} \leftarrow Gaussian(V,\delta 2)$ |
| **8:** | $H_{DOG} \leftarrow (H_{g2} - H_{g1}) \times 200$ , $V_{DOG} \leftarrow (V_{g2} - V_{g1}) \times 200$ |

The results after the horizontal/vertical Gaussian differenced processing are marked by red curves as shown in Figure 6. In Figure 6, to visually and directly see the positions of the horizontal/vertical frame lines in the original images, the red curves are deliberately put on the gray images. The dotted line frame is the effective features region of the horizontal/vertical frame lines, i.e., the positions of the irregular table's horizontal/vertical lines.

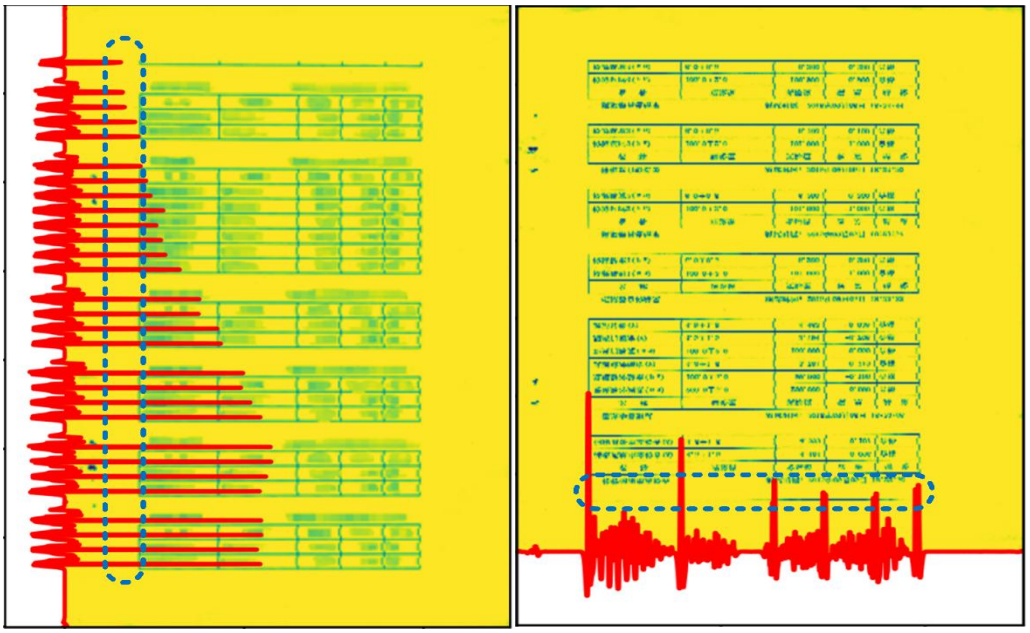


**Figure 6 Results of dual-coding Gaussian difference.**

**Iterative differenced clustering processing method**

This section mainly aims to realize the extraction of table frame line and table boundary frame line. In view of Figure 6, with horizontal $H_{M\times 1}$ and $H_{DOG}$ as examples, the process of iterative differenced clustering processing is introduced. Firstly, the difference of Gaussian results is processed, the time sequence is used to solve the local peaks and troughs, and the iteration is further utilized to process the local peak sequence for obtaining effective table frame lines. Secondly, the differenced feature sequence among the adjacent table frame lines is calculated. Thirdly, the local peaks and troughs sequences in different sequences are solved. Finally, the trough sequence solved by difference is adopted to process the difference peak sequence, and thus the boundary frame lines of the final table are obtained.

The pseudocode of iterative differenced clustering flow is shown in Table 5.

**Table 5**
**Pseudocode of Iterative Differenced Clustering Flow**

| Input: $H_{DOG}$ ; Output: Horizontal effective clustering sequence |
|---|
| 1: Solve local peak sequence and local trough sequence; |
| 2: Iterate the peak sequence according to trough sequence; Mainly aim to solve the effective feature information in blue frame lines; |
| 3: Calculate the effective feature information difference sequence among adjacent differences; |
| 4: Solve the local peak sequence and local trough sequence of difference clustering sequence; |
| 5: Iterate difference peak sequence according to difference trough sequence; |
| 6: Solve the clustering sequence of table horizontal lines according to difference peak **sequence;** |

The output results of iterative difference clustering are shown in Figure 7, in which "+" representsthe local trough, and "." representsthe local peak. In clustering images (c) and (d), the red arrows represent the clustering boundaries, mapping to the table, and the arrows represent the effective clustering boundaries of table frame lines.

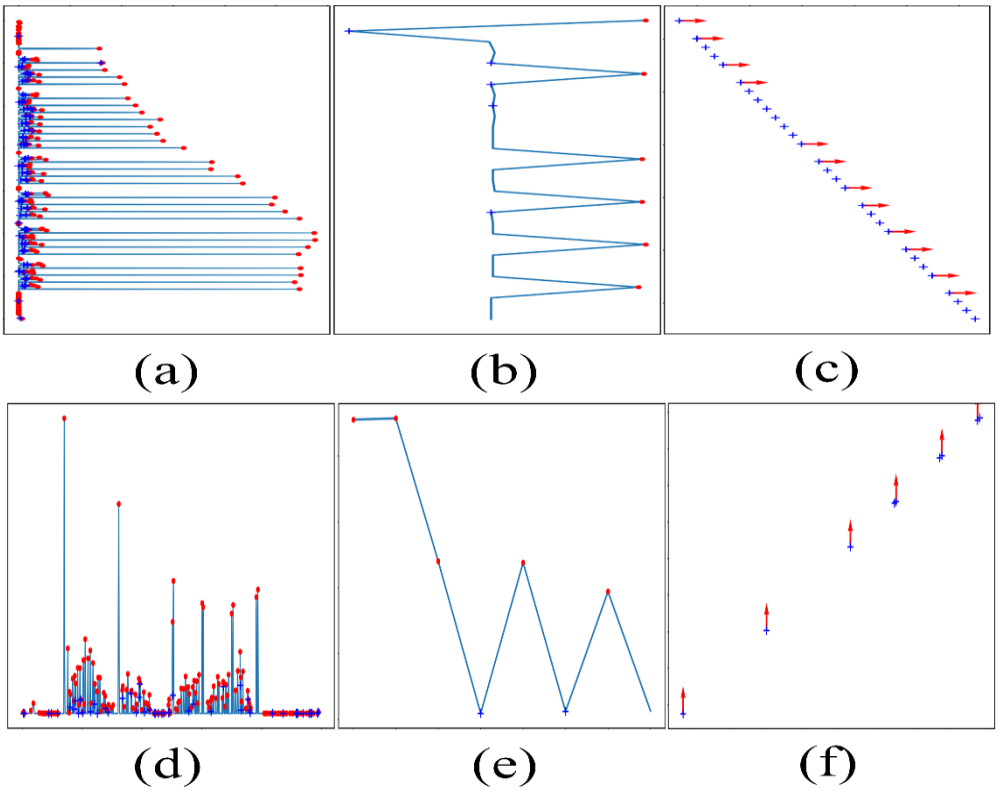



**Figure 7 Results of iterative difference clustering.**

Figures 7(a-c)/(d-f) represent the local peak detection results of horizontal/vertical frame lines, adjacent difference results, and the horizontal frame line clustering results.

**Table reconstruction TSR**

The phenomenon of the multi-page table will occur in the first position and the last position of the table image sequence on the current page. Both may occur in the condition of table frame line number being 1 on the current page. This section takes the multi-page table with an incomplete bottom as an example and introduces the study of the features of the multi-page table in this paper. The incomplete bottoms merely have the following two conditions: this page has incompleteness with content, thus the top of the next page must have incompleteness without content and this page has incompleteness without content, thus the top of the next page must have incompleteness with content.

According to the results of clustering analysis, the table inner structure is reconstructed, as shown in Figure 8.

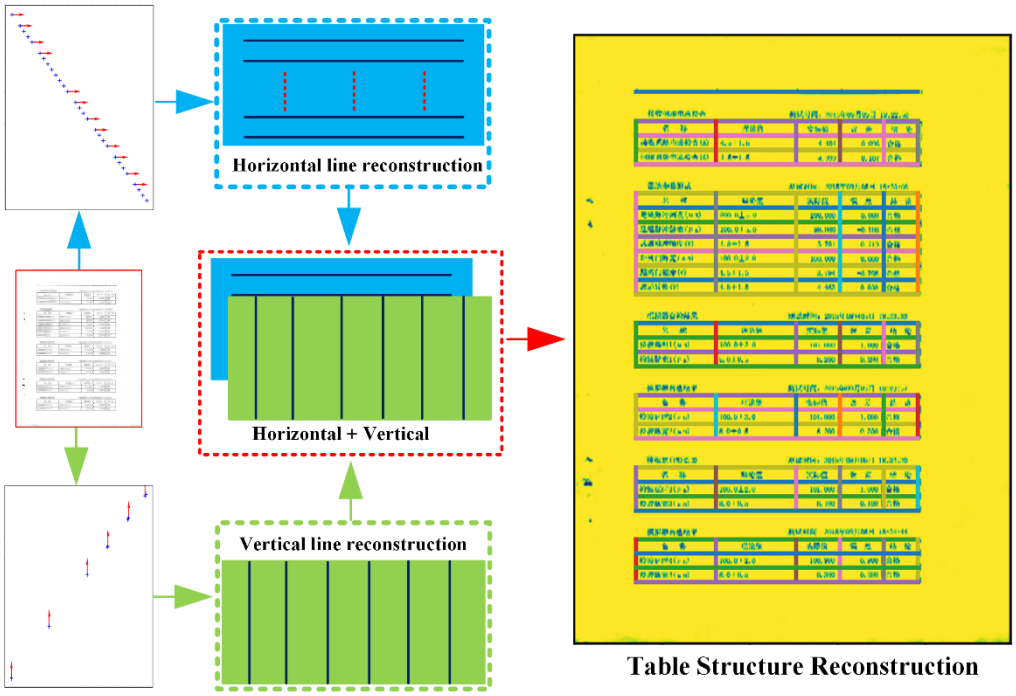



**Figure 8 Reconstruction of table inner structure**

Based onthe output results in Figure 8, the incompleteness can be distinguished by comparing the vertical local pixel features of the incompleteness with and without content. Support vector machine is adopted to perform binary classification of the local pixel features. The classification results can be divided into two states: incompleteness (positive sample) and completeness (negative sample). Equations (12) and (13) define the local regions to be classified. When it is discriminated as an incompleteness state, the vertical frame line is updated to contain the incompleteness area.

Top local positioning region:

$$\begin{aligned} T_x &= V_C(0) - 10 \\ T_y &= H_C(0) - meanrowledge * 0.75 \\ T_w &= V_C(j) + 10 - T_x \\ T_h &= T_y + meanrowledge * 0.5 \end{aligned} \tag{12}$$

Bottom local positioning region:

$$\begin{aligned} B_x &= V_C(0) - 10 \\ B_y &= H_C(i) + meanrowledge * 0.25 \\ B_w &= V_C(j) + 10 - T_x \\ B_h &= T_y + meanrowledge * 0.5 \end{aligned} \tag{13}$$

where, $(T_x, T_y)/(B_x, B_y)$ represent the initial coordinates of top/bottom local regions, $T_w T_h / B_w B_h$ represent the width/height of top/bottom local regions, $(V_C(0), H_C(0))$ represent the position coordinates of the first line after clustering, and $(V_C(j), H_C(i))$ represent the position coordinates of the last line after clustering.

In view of the multi-page table, the traditional image processing method will not be used any longer, and the OCR processing is performed for the existing tables to obtain the table structure. Then, in a combination of recognized text information, the final complete table is reproduced. According to Equations (12) and (13), the positive samples of local regions to be classified are extracted. Figures 9 (a), (b) and (c) show all extracted local region samples and all belong to incomplete positive samples of tables.

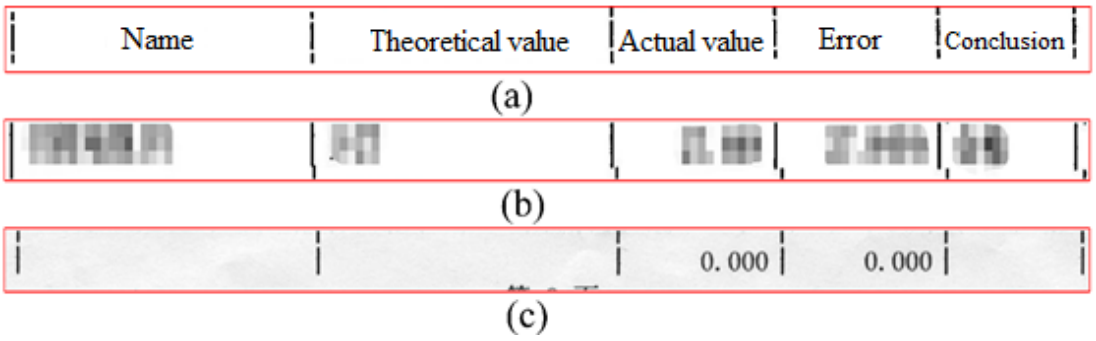



**Figure 9 Positive samples of local regions to be classified**

Through the discrimination for local region classification, it can analyze whether the current table is input into an incomplete multi-page table. The recognized information content is merged to guarantee the data completeness.

**Text information recognition method**

The text information recognition parts are mainly classified into table cell pre-processing, and table cell character recognition. The table cell pre-processing is mainly removing the blank non-text regions. The text images after processing are input to the well-trained character recognition model for performing recognition.

(1) Table cell pre-processing

In the table cells after segmentation, the left and right sides of text have blank areas, which will affect the OCR recognition as well as digital reproduction's accuracy. Hence projection needs to be used to make further processing. The table cell images to be processed are input. Firstly, the up and down blank areas of the text region are removed through horizontal projection statistics. Secondly, the left and right blank areas of the text are removed through vertical projection statistics. Figure 10 shows the boundary processing of the table cells. Figure 10(a) is the cell image of the input positioning area. Figure 10 (b) is obtained by adopting horizontal projection statistics for Figure 10(a) and removing the blank areas. Figure 10 (d) is obtained by adopting the vertical projection statistics for Figure 10 (b) and removing the blank areas. Next, the construction of a digital reproduction recognition model for the images after processingis introduced.

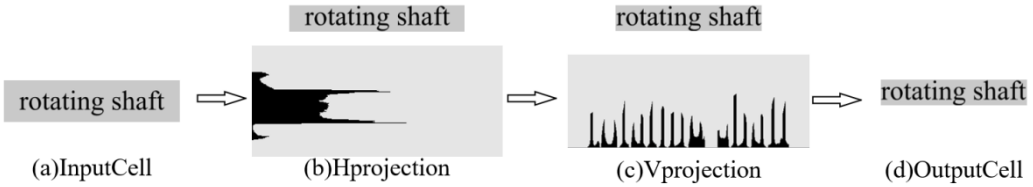



**Figure 10 Processing of table cell boundary**

(2) Character recognition of table cell

Convolutional recurrent neural network model (CRNN) of end-to-end is selected as the network model of OCR character recognition. Network training sets mainly come from the following three parts: ICPR MTWI 2018 dataset, the dataset generated by program simulation, and the dataset obtained by practical sample cutting.

The network model structure table based on the CRNN model is shown in Table 6, which adopts the Python3+Tensorflow framework to realize the model's design and training.

**Table 6**
**Network Structure**

| Type | Configurations |
|---|---|
| Input | W×32 gray-scale image |
| Convolution | maps:64,k:3×3,s:1,p:1 |
| MaxPooling | window:2×2,s:2 |
| Convolution | maps:128,k:3×3,s:1,p:1 |
| MaxPooling | window:2×2,s:2 |
| Convolution | maps:256,k:3×3,s:1,p:1 |
| Convolution | maps:256,k:3×3,s:1,p:1 |
| MaxPooling | window:1×2,s:2 |
| Convolution | maps:512,k:3×3,s:1,p:1 |
| BathNormalization | - |
| Convolution | maps:512,k:3×3,s:1,p:1 |
| BathNormalization | - |
| MaxPooling | window:1×2,s:2 |
| Convolution | maps:512,k:2×2,s:1,p:0 |
| Map-to-Sequence | - |
| Bidirectional-LSTM | hidden units:256 |
| Bidirectional-LSTM | hidden units:256 |
| Transcription | - |

Hardware configurations used in the training are shown in Table 7

.

**Table 7**
**Hardware Configurations**

| Name | Device parameters |
|---|---|
| CPU | Intel(R) Xeon(R) W-2123 |
| Memory | 32G |
| Operating system | Windows10 |
| GPU | NVIDIA GeForce RTX 2080Ti |
| Debugging environment | PyCharm |
| Deep learning framework | Tensorflow1.13.1 |

Before training, data enhancement for images was performed by overturning, rotating, zooming, cutting, and adding noise in order to improve the accuracy of the model. During training, the network model with the parameter configurations in Table 8 was adopted. Through iterative training for 100 Epochs, the well-trained CRNN modelwas obtained and saved with checkpoint format for consequent recognition and call.

**Table 8**
**Configurations of Training Parameters**

| Name | Model parameters |
|---|---|
| batch_size | 64 |
| Initial learning rate | 0.001 |
| Damping cycle of learning rate | 400 |
| Damping coefficient of learning rate | 0.47 |
| Optimizer | Adam |
| Epoch | 100 |

After 2500 iterations, Tensorboard was used to realize model visualization and obtain the loss of iterative training, as shown in Figure 11

.

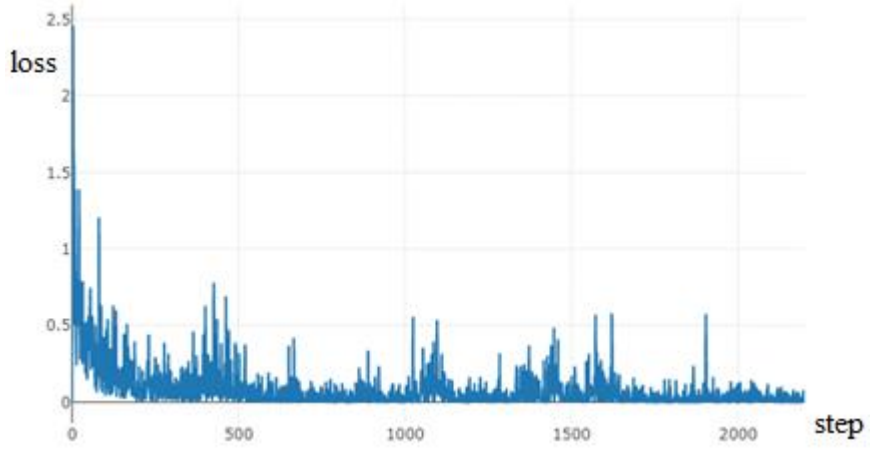



**Figure 11 Loss curve of iterative training.**

The model is tested in 10373 images, with recognition accuracy reaching above 95%. The recognition results are shown in Figure 12. In Figure12, the ordinate valve represents the accuracy rate of predicting, 0.8 representing 80% for example.

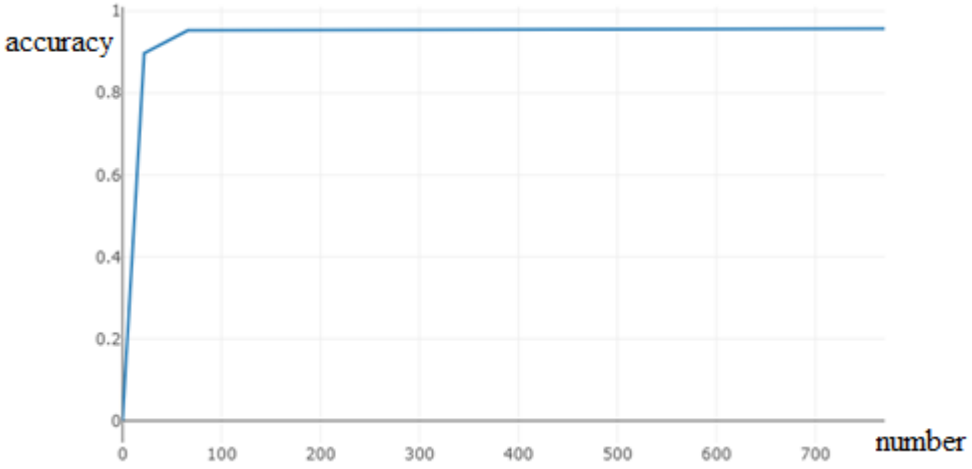


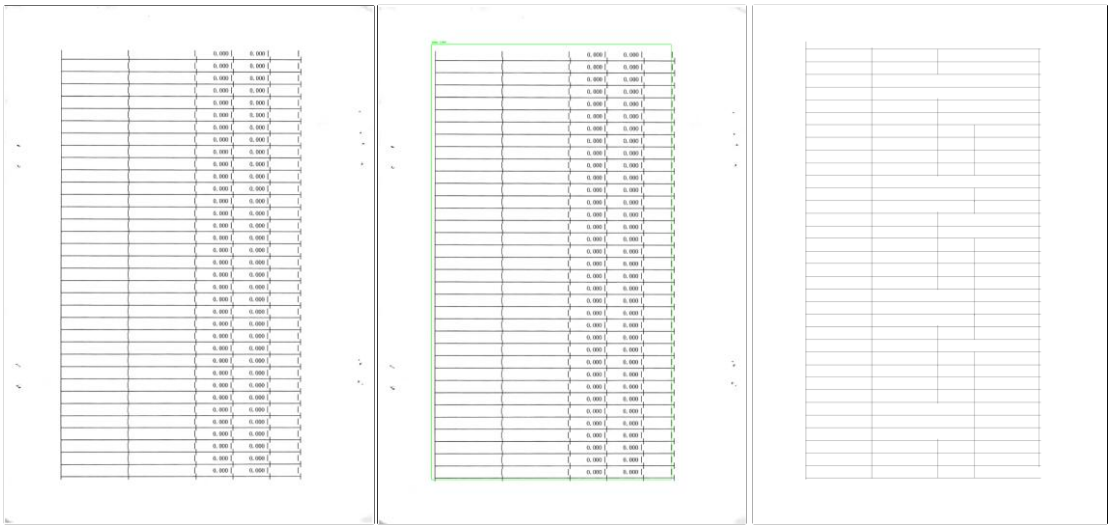

**Figure 12 The accuracy of the model on the test set**

## EXPERIMENT AND ANALYSIS

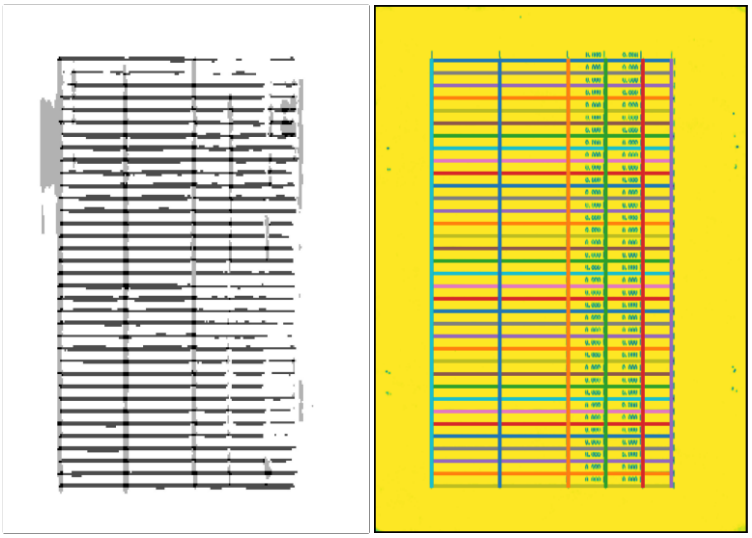
The algorithm was implemented on Windows 10 platform, with Intel (R) Core (TM) CPU i7-8700HQ 3.20GHz 8G memory. The scanning equipment was Alaris E1025 and EPSON Perfect v19. For the non-deep learning part, C++ programming language was adopted, and the program realization was conducted on QT5.9.8, MSVC2015 compiler and Opencv3.1. At the same time, for the deep learning part, Python3.6 integrated development environment was partly adopted and the CRNN deep learning model was constructed on Tensorflow1.13.1 edition forrealizing the OCR recognition.

### Qualitative analysis

Under the same experimental conditions, in terms of table images with different resolutions, the algorithm in this paper, Mask-RCNN, UNET, YOLOV3+UNET, AlexNet improved by deconvolutional up-sampling, and TableParser-TX, areused respectively to perform table positioning as well as the reconstruction of the table inner structure. Some comparative test results are shown in Figure 13.



(a)Fuzzy input sample (b)Mask-RCNN (c)UNET



(d) TableParser-TX (e)OURS

**Figure 13 Somecomparative test results**

The experimental results show that in randomly extracted test samples, the algorithm proposed in this paper can realize accurate positioning segmentation. The UNET and Mask-RCNN cannot realize positioning segmentation well for some special tables. The UNET and YOLOV3+UNET have poor reconstruction effects of inner structure in some tables. When the frame lines are critically misplaced, the TableParser-XT proposed by Tencent cannot realize good frame line reconstruction. In the experiment of adopting the AlexNet improved by deconvolutional up-sampling, the yellow part is the table region, and the other colors are all non-table regions. It can be analyzed from the experimental results that these regions do not obtain positive classification detections. For the irregular tables with the phenomenon of discontinuous table frame lines or critically misplaced frame lines, the use of the dual-coding difference of Gaussian processing algorithm proposed in this paper can well realize the table positioning as well as the reconstruction of inner structure.

**Quantitative analysis**

For the existing sample set consisting of 12,840 table images, typical samples were randomly selected. The test results in Figure 13 indicate that the algorithms above have many problems while processing the irregular tables. During the process of quantitative analysis, the recognition rate usually serves as an evaluation indicator.

The recognition rate is defined as follows:

$$R = (\frac{Correct\_identification}{All\_identification}) * 100\% \quad (14)$$

For the sample set consisting of multi-resolution table images after skew correction, iteration tests were conducted for the positioning accuracyand the structure reconstruction accuracy of tables. The test results of average recognition accuracies are shown in Table 9.

**Table 9**

**Test Results of Average Recognition Accuracies**

| Input | Method | Table/ R% | TableCell/R % | Integrality/R % |
|---|---|---|---|---|
| Vertical broken | Mask-RCNN | 0.959 | None | None |
| | UNET | 0.985 | 0.965 | None |

| | | | | |
|---|---|---|---|---|
| line | YOLOV3+UNET | 0.853 | 0.445 | None |
| | AlexNet | 0.671 | None | None |
| | TableParser-TX | **1.000** | 0.232 | None |
| | OURS | **1.000** | **0.996** | **0.968** |
| Vertical malposition line | Mask-RCNN | 0.948 | None | None |
| | UNET | 0.974 | 0.971 | None |
| | YOLOV3+UNET | 0.827 | 0.249 | None |
| | AlexNet | 0.559 | None | None |
| | TableParser-TX | **1.000** | 0.252 | None |
| | OURS | 0.979 | **1.000** | **0.949** |

The results in Table 9 indicate that the TableParser-TX algorithm can accurately position the table region, the table's horizontal frame line reconstruction effects are rather good, and the horizontal frame lines of all tables can be extracted. However, the reconstruction effect of the vertical frame line's dotted lines or misplacedtable structure is not good. The UNET has a rather good effect in terms of table positioning and table inner structure reconstruction, but has a poor effect in terms of table inner processing with frame line's misplaced part. Therefore, it should be further studied in depth. If the vertical frame lines with specified features can be accurately segmented and extracted, the accurate reconstruction of the table's vertical frame lines can be realized theoretically. From the qualitative and quantitative results, the UNET has good effects in aspects of the positioning and structure reconstruction of irregular tables. Relatively, the irregular table processing algorithm proposed in this paper has certain practical feasibility, but the excellence of algorithm implementation effect depends on the effect of correction. Under the condition of poor correction effect, misjudgment against table positioning and inner structure reconstruction will arise.

## CONCLUSION

Considering the current problems that the OCR algorithm cannot recognize irregular tables and cannot realize the relevance of context information for multi-page tables, this paper proposes an irregular table image processing algorithm based on the dual-coding difference of Gaussian iterative clustering. Artificial intelligence methods (e.g., SVM, CRNN) are successfully combined with the traditional morphological processing methods to solve the recognition problems of table elements and character elements in regular/irregular tables. It can be seen from the qualitative and quantitative analysis results that the algorithm proposed inthis paper has higher university and recognition accuracy,and is suitable for engineering implementation, compared with the other OCR algorithms.

## Acknowledgment

This study is supported by the following funds: Key Research & Development Project, Science and Technology Department of Shaanxi Province (No. 2019GY-065); Xi'an Science and Technology Planning Project (No.2020KJRC0033); and Science and Technology Planning Project of Weiyang District, Xi'an City (No.201923).

## References

1. Kuang Z,Cui Z.Form recognition algorithm in community election system. Computer applications. 2017;37(S2):179-182.(in Chinese).
2. DuanL, Song YH, Zhang YL. A layout analysis algorithm for questionnaire image. *Journal of software*. 2017;28 (02): 234-245. (in Chinese) .
3. Bai W, Cui Z. Table frame-line detection algorithm based on run-distance clustering. *Computer applications*. 2008;38(S1):179-182.(in Chinese).
4. Bansal A, Harit G, Roy S D, et al. Table Extraction from Document Images using Fixed Point Model. *Proceedings*

*of the 2014 Indian Conference on Computer Vision Graphics and Image Processing.ACM*. 2014:67

5. Ohta M, Yamada R, Kanazawa T, et al. A Cell-detection-based Table-structure Recognition Method. *Document Engineering*. 2019: 3345412

6. Xiao D, Sun HY, BaoZL. A Multi-Table Image Recognition System Based on Deep Learning and Edge Detection. *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Hubei Zhongke Institute of Geology and Environment Technology*、*Liverpool John Moores University. UK*.(AICS 2019);2019:200-207.

7. Liang QK,Peng JZ,Li Z W,etal.Robust table recognition for printed document images.*Mathematical Biosciences and Engineering (MBE)*. 2020; 17(4): 3203-3223.

8. Kim J, Hwang H. A Rule-based Method for Table Detection in Website Images.*IEEE Access*. 2020; PP(99):1-1.

9. Smita P,Raj Ratn P,Sumit K. A Conglomerate of Multiple OCR Table Detection and Extraction.*2020 International Conference on Document Analysis and Recognition(ICDAR)*.2020.

10. Shahzad M A, Noor R, Ahmad S, et al. Feature Engineering Meets Deep Learning: A Case Study on Table Detection in Documents. *Digital image computing: techniques and applications*. 2019:1-6.

11. Li YB,Yan QQ,Huang YL,et al.A GAN-based Feature Generator for Table Detection. *2019 International Conference on Document Analysis and Recognition(ICDAR)*. 2019:763-768.

12. Riba P, Dutta A, Goldmann L, et al. Table Detection in Invoice Documents by Graph Neural Networks.*2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019.

13. Paliwal S, Vishwanath D, Rahul R, et al. TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images. *International Conference on Document Analysis and Recognition (ICDAR)*. 2019:128-133.

14. Huang YL, Yan QQ, Li YB, et al. A YOLO-Based Table Detection Method.*International conference on document analysis and recognition*. 2019.

15. Kavasidis I, Palazzo S, Spampinato C,et al.A Saliency-based Convolutional Neural Network for Table and Chart Detection in Digitized Documents. *International conference on image analysis and processing*. 2019: 292-302.

16. Prasad D, Gadpal A, Kapadni K, et al. CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.2020:572-573.