Triangle Coordinate Diagram Localization for Academic Literature Based on Line Segment Detection in Cloud Computing

Baixuan Tang[1], Jielin Jiang[1], Xiaolong Xu[1(✉)], Lianyong Qi[2], Xiaokang Zhou[3], and Yang Chen[4]

[1] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
njuxlxu@gmail.com

[2] School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China

[3] Faculty of Data Science, Shiga University, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

[4] School of Computer Science, Fudan University, Shanghai, China

Abstract. The localization of triangle coordinate diagram in academic literature is an important step in the process of data mining. However, the detection of triangle coordinate diagram in academic literature mainly depends on manual work, which consumes a lot of time. At present, there is no specific locating method for triangle coordinate diagram. To solve this problem, this paper proposes a method of triangle coordinate diagram localization based on line segment detection, which can be placed on the cloud platform to provide convenience for the functions such as diagram-based retrieval in academic literature database. Technically, this paper uses line segment detection algorithm and line segment merging algorithm to complete triangle coordinate diagram localization. Finally, an experiment is conducted to evaluate the method, which proves the effectiveness of the proposed method.

Keywords: Triangle coordinate diagram · Line segment detection · Academic literature · Diagram localization

1 Introduction

Diagram is an important part of academic literature. As a special non-text data structure in literature, it can reflect the key research methods and conclusions in a great measure. With the explosion of academic literature in recent years, the visual resources in the literature present an explosive growth trend. Database vendors continue to strengthen the disclosure of visual resources in the literature, add diagram-based retrieval, build public knowledge resource pools based on academic visual resources, and carry out related services in the search and open access of resources [1]. At the same time, the academic literature database based on the cloud platform makes it possible to place algorithms

M. R. Khosravi et al. (Eds.): CloudComp 2021, LNICST 430, pp. 47–59, 2022.
https://doi.org/10.1007/978-3-030-99191-3_5

with high requirements on computer performance on the cloud, so that these algorithms can be conveniently applied in more common scenarios. Diagram can also help readers quickly and directly understand the author's intention and the core idea of the literature, which is of great help to readers. Scholars try to help readers by analyzing the diagram in the literature, such as searching and recommending appropriate literature. It can be predicted that the development and construction of visual resources in the paper will further promote the organization and dissemination of academic knowledge and have a broad development space in the field of knowledge service in the future.

Compared with the text content in the literature, there are relatively few studies on the diagrams. One of the main reasons is that in the common academic storage format, PDF, the content is not logically structured, but simply positioned by using localization. In other words, in PDF format, diagrams are stored in a discrete manner, and the elements within the diagrams are structurally independent, so they cannot be located directly from the academic literature.

Nowadays, common research on the localization of diagrams in academic literature is mainly based on the underlying coding of PDF, and the localization of specific types of diagrams according to the prior rules of typesetting and literature writing. To some extent, it has a beneficial effect for the same type of literature in PDF format, which is relatively fixed in typesetting format. However, because it mainly analyses the underlying coding of PDF, it ignores the visual information on the page. Nowadays, there are many kinds of academic literatures, which is difficult to extract information accurately with limited typography rules. At the same time, in reality, a large number of electronic literature data of early literature are stored in the form of image after scanned by paper literature, and diagram localization based on the PDF format is more difficult to apply in this case.

Triangle coordinate diagram usually use three-dimensional percentage coordinates to represent the proportion of an element in a three-component system to the overall structure. As a kind of structural information image, three sides of trigonometry represent three different elements, three vertices represent three origins, and the three-dimensional coordinates of points represent the proportion of each component in an element.

Triangle coordinate diagram can be used as a classification diagram by classifying each element according to its component proportion. Triangle coordinate diagram is widely used in geology, chemistry and statistics, such as the classification of sandstone and greywacke, classification of soil texture, chemical classification of carbonate, representation of population age structure and so on.

Different from the common diagram with rectangle as the border, triangle coordinate diagram is a kind of diagram with a triangle as the border, which makes the locating method of this diagram different from the locating method of common diagram. At present, there is no diagram localization algorithm specifically for triangle coordinate diagram. The method proposed can be combined with diagram retrieval and participate in the work related to triangle coordinate diagram data extraction to save manpower and improve efficiency. Considering the wide application of triangle coordinate diagram in all kinds of literature, it has certain research value to study the localization of triangle coordinate diagram in academic literature.

According to the geometric characteristics of triangle coordinate diagram in literature, a method of triangle coordinate diagram localization based on line segment

detection is proposed. At the same time, this method does not require a lot of data annotation, which can save a lot of manpower. Specifically, the contributions of this paper are as follows:

1. A line segment merging method is proposed, which can further process the line segments after line segment detection. This method can merge the segments that are incorrectly segmented and the segments that are repeatedly recognized.
2. A method to locate the triangle coordinate diagram according to the detected line segments is proposed.

The rest of this article is organized as follows: The related work in this article is described in Sect. 2. In Sect. 3, the methods of academic literature triangle coordinate diagram localization are described. Experiment are conducted in Sect. 4. In Sect. 5, we conclude this article and look forward to the future.

2 Related Work

Diagram is an important part of literature, which can express the information that needs complex text description in an intuitive and concise way, so that readers can better understand what the author wants to express [2]. Diagram contains the core content of literature, which has high research value and is an important research object. Lee et al. [3] used machine learning technology to divide 8 million diagrams into five categories according to the content. Through the comparison of their influence with the corresponding literature, it was found that the distribution and types of diagrams remained relatively stable for a long time, but there were differences in different fields. Among them, the higher the influence of literature, it tends to use more schematic diagrams to help readers understand. Apostolova et al. [4] pointed out that the accurate localization and indexing of images in literature has an important impact on the accuracy and efficiency of literature image retrieval.

From the above research, it is not difficult to see that the academic literature diagram has high research value. The research on diagrams in academic literature can be divided into academic diagram retrieval [5], diagram similarity calculation [6], and diagram content analysis and extraction [7]. These researches need to accurately locate the diagrams in the literature before they can be carried out. It can be seen that the accuracy and comprehensiveness of the diagram localization in the literature have a great impact on the follow-up research.

There are two kinds of research objects on the diagram localization of academic literature: Scanned academic literature (including electronic literature in the form of pure pictures, most of which are stored in the form of pictures) and literatures in PDF format (most of the text content is stored in the form of characters). For the method of scanned literature, because the main content of the literature is stored in pictures, the academia treats the problem as a kind of image processing problem to solve. Specifically, this method can be roughly divided into two types, one of which divides the picture from the whole page from top to bottom to get the final diagram area, and then classifies the area to locate the existing diagram in the literature in the picture [8]; Another method starts

with pixels, calculates the connectivity between the contents from bottom to top, merges the pixels according to certain rules, and finally merges them into a complete diagram area image [9]. For the PDF format literature method, using the format characteristics of the file, extract specific structure content from it according to certain rules, and locate the diagram by matching keywords and searching for areas without body content near keywords. For example, PDF Figures [10] literature diagram localization tool, published by Allen AI Lab, University of Washington, uses heuristic algorithms to accomplish the task of graph positioning by means of label positioning, area content recognition, and performs well on its own dataset.

In recent years, there have been a large number of methods based on deep learning in academia on the localization of diagram and data fetch in the literature. Ma et al. [11] uses deep neural network to extract the semantics of scatterplots in academic papers, and completes the task of calculating the similarity of two scatterplots. Yu et al. [12] extracted the diagrams from the academic papers on artificial intelligence, classified them and located the diagrams describing the deep learning model, then used these diagrams to generate standard flowcharts and get the corresponding codes. The deep learning method can automatically extract the characteristic information from the dataset, and it performs well in the field of computer vision, such as image classification, target detection, and so on. It surpasses human performance in some tasks [13]. However, at the same time, some studies [14] have shown that the lack of high-quality training data limits the ability of deep learning to solve problems on related tasks. There is a large demand for data to solve problems using deep learning, which also greatly increases the cost of research to solve related problems.

Summarizing previous studies, the following areas need to be improved:

(1) The electronic document encoding based diagram localization method for PDF can only handle documents in PDF format, but cannot handle the problem of diagram localization in a large number of scanned documents (stored as picture format); (2) The method based on deep learning requires manual labeling of a large number of data for the extracted diagram target, which makes it difficult to cope with the lack of related data.

According to the characteristics of triangle coordinate diagram in academic literature, this paper proposes a method of triangle coordinate diagram localization in academic literature based on line segment detection: (1) The literature image is de-noised by Gaussian filter to get low noise image, and the image edge is detected by Sobel operator to obtain the binary image; (2) The original line segment set is obtained by line segment detection of low noise image; (3) Processing the original line segment set: merging the overlapping and wrong segments to get the merged line segment set; (4) According to the relative position between each line segment in the triangle segment set, whether the triangle segment group is a triangle is judged.

3 Triangle Coordinate Diagram Localization

3.1 Gaussian Filter and Sobel Operator

The literature image is de-noised to reduce the redundant noise information in the image, which provides convenience for the next step of line segment detection.

According to the values of red channel R, green channel G and blue channel B of the image, the original image is converted into gray image I by using gray calculation formula.

The gray calculation formula is as follows:

$$\boldsymbol{I} = \boldsymbol{R} \times 0.299 + \boldsymbol{G} \times 0.587 + \boldsymbol{B} \times 0.114 \tag{1}$$

when all the pixels of the image are transformed into gray image by gray operation, Gaussian blur formula is used to convolute the gray image I and Gaussian kernel.

The Gaussian blur formula is as follows:

$$\boldsymbol{I}_{\sigma} = \boldsymbol{I} * Gaussian_{\sigma} \tag{2}$$

where * denotes convolution operation, $Gaussian_{\sigma}$ is a two-dimensional Gaussian kernel with a standard deviation of σ, which is defined as:

$$Gaussian_{\sigma} = \frac{1}{2\pi\sigma} \mathrm{e}^{-(x^2+y^2)/2\sigma^2} \tag{3}$$

where x and y represent the abscissa and ordinate of the pixel respectively; After the whole image is convoluted by Gaussian kernel, the de-noised image is obtained.

In order to better detect the line segments in the image in the next step, Sobel operator filtering is also needed for the image.

Sobel operator contains two convolution kernels, namely transverse convolution kernel and longitudinal convolution kernel. The biggest difference between triangle coordinate diagram and common diagram is that there are segments with an inclination of about 60°. This paper uses these inclined segments as the main basis to judge whether there is a triangle coordinate diagram in the image and locate the triangle coordinate diagram. In practice, there are a large number of transverse straight lines in literature images. Therefore, Sobel transverse convolution kernel is used here to convolute the image. In this way, we can get the image y after filtering out the transverse edge. The formula is as follows:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * y \tag{4}$$

where $*$ denotes convolution operation, G_x represents the image gradient value of transverse edge detection.

3.2 Line Segment Detector

In this step, line segment detection will be performed on the literature image to obtain the line segment set. Specifically, LSD [15] is used to detect line segment.

The main process is as follows:

(1) Calculate the gradient angle θ and gradient G of each pixel in the image, the formula is:

$$\theta(x, y) = \arctan\left(\frac{g_x(x, y)}{-g_y(x, y)}\right) \tag{5}$$

$$G(x, y) = \sqrt{g_x^2(x, y) + g_y^2(x, y)} \tag{6}$$

where g_x and g_y represents the horizontal and vertical gradient values of the pixel respectively, the formula is:

$$\begin{aligned} g_x(x, y) &= \frac{i(x+1,y)+i(x+1,y+1)-i(x,y)-i(x,y+1)}{2} \\ g_y(x, y) &= \frac{i(x,y+1)+i(x+1,y+1)-i(x,y)-i(x+1,y)}{2} \end{aligned} \tag{7}$$

where $i(x, y)$ is the gray value of the pixel at (x, y);

(2) The direction of the pixel is combined into the direction of the line area, and the pixel is filtered by judging the direction of the pixel and the direction of the line area. Select an unselected pixel as the seed point to judge the pixel: The other pixels whose difference between gradient angle and region angle is less than the threshold is added to the line support domain. Every time a new pixel is added to the area, the region angle of the whole line area is update. The angle formula of an area is as follows:

$$\arctan\left(\frac{\sum_j \sin\theta_j}{\sum_j \cos\theta_j}\right) \tag{8}$$

(3) The diffused region is fitted by rectangle, and an external rectangle containing all pixels in the region is constructed. The main axis angle of the rectangle is calculated, and the main axis angle is set as the angle of the line segment to be extracted;

(4) Verify the line segment and detect the rectangle r. According to the corresponding NFA formula, calculate whether the rectangle r meets the threshold: if not, ignore it; if yes, express the rectangle record as a detected line. The formula is as follows:

$$\mathrm{NFA}(r) = (NM)^{5/2} \cdot B(n, p) \tag{9}$$

where N and M represent the column width and row width of the image, and the formula of $B(n, p)$ is:

$$B(n, p) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} \tag{10}$$

where n is the total number of pixels in the rectangle and p is the precision;

(5) Go back to step (2), find the next seed point, spread the rest of the image until traversing the whole image, and get the original line segment set of the image.

3.3 Line Segment Merging

LSD has excellent ability to detect line segments. However, a single line segment in the image of literature may be detected as multiple line segments after line segment detection, which will have a negative impact on the locating and type determination of the diagram. Therefore, after line segment detection, it is necessary to merge the

line segments according to certain rules: delete the redundant segments and connect the segments that are wrongly segmented.

In this paper, according to the specific situation of the literature, some thresholds are preset. When using this method to identify the diagrams in other literature, the thresholds need to be adjusted according to the situation.

(1) First of all, set the line segment end merging threshold and slope merging threshold. The threshold value can be set according to the page size of literature image, or can be manually specified according to experience;

(2) Then, the redundant segments are merged. Firstly, the redundant line segments with similar length and position are merged. The method is as follows:

Judge whether there are two pairs of endpoints in two line segments, and the distance between the two pairs of endpoints is less than the endpoint merging threshold. If there is a group of line segments that meet the conditions, merge them: According to the formula, the center point of each pair of end points is taken as the end point of the merged line segment, a merged line segment is generated, and two original line segments are deleted. The end coordinates (X, Y) formula of merging line segments is as follows:

$$X = \frac{x_0 + x_1}{2} \tag{11}$$

$$Y = \frac{y_0 + y_1}{2} \tag{12}$$

where (x_0, y_0) and (x_1, y_1) are the two endpoints to be merged;

After that, merge the redundant segments with length difference. The method are as follows: Judge whether the slope difference of two line segments is less than the slope threshold, and whether the maximum distance between the two ends of the shorter line segment and the longer line segment is less than the point line merging threshold.

If the conditions are met, the segments are merged: The shorter segments are deleted, and the longer segments are retained;

(3) Finally, the segments that are wrongly segmented are merged. For the wrong segment, the method is as follows: The distance between two segments is less than the merging threshold, and the slope difference between the two segments is less than the slope threshold.

If the conditions are met, merge segments: Take the two farthest endpoints of two line segments as the two endpoints of the merging line segment, generate a merging line segment, and delete the two original line segments;

(4) Repeat steps (2) and (3) for the line segment set until the set is no longer updated, which means that the line segment merging is completed.

3.4 Diagram Localization

After the above steps, the line segments in the set can represent the frame of the triangle in the literature image to a certain extent. Next, according to certain steps, find the appropriate line segment pair from the set for matching. The steps are as follows:

(1) Classify the line segments in the set according to the slope. Line segments with positive slope form set 1; Line segments with negative slope form set 2.

(2) Select a line segment from set 1. In set 2, select a segment whose absolute value of slope is close to it and whose endpoint is closest to the endpoint of the selected segment in set 1.
(3) Record the selected line segments in set 1 and set 2, and remove the two segments in set 1 and set 2. The two line segments are regarded as the waist of the triangle coordinate diagram in the literature image. Each pair of such line segments locates a triangle coordinate diagram.
(4) Repeat steps (2) to (3) until all line segments in set 1 cannot find qualified line segments in set 2, or one of the sets is empty.

The algorithm from step (2) to step (3) is as follows:

Algorithm 1 matchLine(set1, set2)

```
for each line1 in set1 do
    for each line2 in set2 do
        if (|slope of line1 – slope of line2| < threshold)
        and (endpointDistance(line1, line2) < minDistance)
        then
            minDistance := endpointDistance(line1, line2)
            targetLine := line2
        end if
    end for
    add (line1, targetLine) to targetSet
end for
```

where function endpoint Distance(line1, line2) calculates and returns the minimum distance between two endpoints of line1 and two endpoints of line2.

This paper combines LSD line segment detection [15] and line segment merging to establish a new triangle coordinate diagram localization method. Compared with the common line segment detection methods, we add the line segment merging step after the line segment detection, which is very helpful to judge the diagram type according to the spatial relationship of line segments.

4 Experiment

4.1 Dataset

In order to verify the performance of the proposed method, this paper uses the self-built academic literature triangle coordinate diagram dataset for experiments. We searched the open academic literature dataset for keywords such as “decimal composition”, “sandstones”, and “provenance”. After manual judgment, 43 literatures containing triangle coordinate diagram were selected and constructed into the self-built dataset of this paper. The dataset contains 43 academic literature pages in the format of pictures, in which

each page contains at least one triangle coordinate diagram, and all pages contain 112 triangle coordinate diagrams.

4.2 Experimental Method

An example of a literature image in the dataset (Fig. 1) shows how this method locates triangle coordinate diagrams.

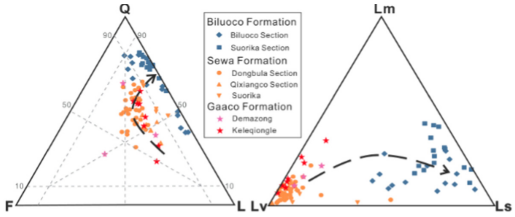

Figure 6. Ternary diagrams for sandstones of the Sewa, Gaaco, and Biluoco formations. Q, quartz; F, feldspar; L, lithic fragments (Lm, metamorphic; Ls, sedimentary; Lv, volcanic). The black arrows highlight the compositional change consequent to the late Bathonian tectonic event.

Sandstone compositions are different in the north Biluoco and Suorika sections of the Abushan Formation. In the north Biluoco section, sandstones of the Abushan Formation are litho-quartzose and quartzo-lithic volcaniclastic (average composition QFL = 46:9:45, LmLsLv = 3:24:73) (Table S2). Intermediate to felsic volcanic rock fragments prevail over limestone and minor sandstone clasts. In the Suorika area, sandstone samples are litho-quartzose and quartzo-lithic sedimentaclastic (average composition QFL = 51:47:2, LmLsLv = 8:89:2) (Table S2), with dominant limestone and only minor sandstone and volcanic rock fragments.

Fig. 1. An image in the dataset

(1) Firstly, Sobel operator and Gaussian filter are performed on the literature image to get Fig. 2.

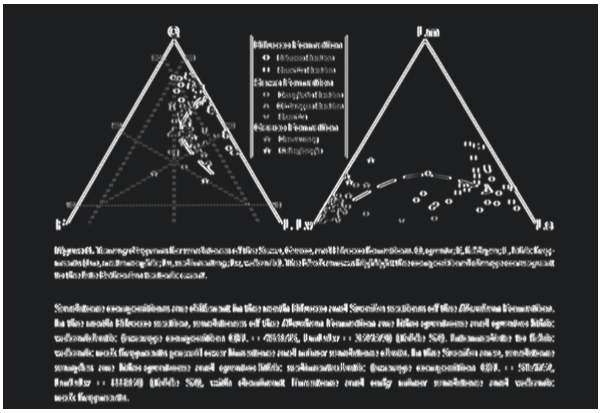

Fig. 2. The image after performed

(2) Line segment detection is performed on the image from the previous step to get the set of line segments. Draw all line segments in the set on literature image, as shown in Fig. 3. (Where each line segment is drawn in a random color, the same below).

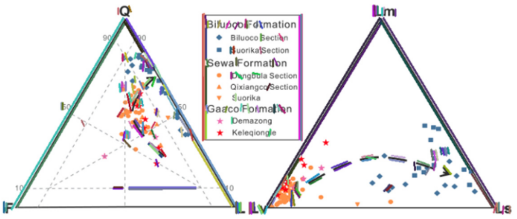

Figure 6. Ternary diagrams for sandstones of the Sewa, Gaaco, and Biluoco formations. Q, quartz; F, feldspar; L, lithic fragments (Lm, metamorphic; Ls, sedimentary; Lv, volcanic). The black arrows highlight the compositional change consequent to the late Bathonian tectonic event.

Sandstone compositions are different in the north Biluoco and Suorika sections of the Abushan Formation. In the north Biluoco section, sandstones of the Abushan Formation are litho-quartzose and quartzo-lithic volcaniclastic (average composition QFL = 46:9:45, LmLsLv = 3:24:73) (Table S2). Intermediate to felsic volcanic rock fragments prevail over limestone and minor sandstone clasts. In the Suorika area, sandstone samples are litho-quartzose and quartzo-lithic sedimentaclastic (average composition QFL = 51:47:2, LmLsLv = 8:89:2) (Table S2), with dominant limestone and only minor sandstone and volcanic rock fragments.

Fig. 3. The image drawn with line segments in the set

(3) Filter out the line segments whose length and angle do not meet the rules in the line segment set of the previous step. Draw the filtered line segment set on the literature image, as shown in Fig. 4. (End points of line segments are marked with red dots).

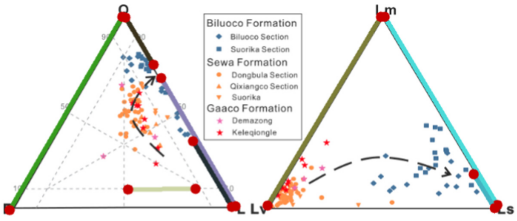

Figure 6. Ternary diagrams for sandstones of the Sewa, Gaaco, and Biluoco formations. Q, quartz; F, feldspar; L, lithic fragments (Lm, metamorphic; Ls, sedimentary; Lv, volcanic). The black arrows highlight the compositional change consequent to the late Bathonian tectonic event.

Sandstone compositions are different in the north Biluoco and Suorika sections of the Abushan Formation. In the north Biluoco section, sandstones of the Abushan Formation are litho-quartzose and quartzo-lithic volcaniclastic (average composition QFL = 46:9:45, LmLsLv = 3:24:73) (Table S2). Intermediate to felsic volcanic rock fragments prevail over limestone and minor sandstone clasts. In the Suorika area, sandstone samples are litho-quartzose and quartzo-lithic sedimentaclastic (average composition QFL = 51:47:2, LmLsLv = 8:89:2) (Table S2), with dominant limestone and only minor sandstone and volcanic rock fragments.

Fig. 4. The image drawn with the filtered line segments

(4) For the line segment set in the previous step, the line segments are merged according to the method proposed in this paper. After that, the line segment set is drawn on the literature image, as shown in Fig. 5.

(5) Search and match the most appropriate line segment from the set in the previous step, and set them as a triangle segment group. Each triangle segment group represents a triangle coordinate diagram, as shown in Fig. 6. (Different triangle coordinate diagrams are represented by triangle boxes of different colors.)

The experiment uses the method in this paper to locate the triangle coordinate diagram in the dataset. The results are given as a triangle box on the rendered page. If no triangle coordinate diagram is found, no box is given in the corresponding area. The result of the

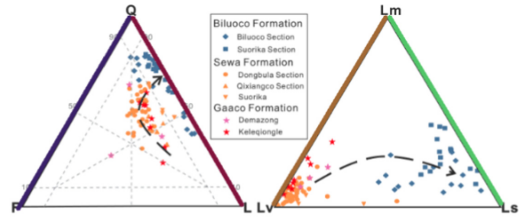

Figure 6. Ternary diagrams for sandstones of the Sewa, Gaaco, and Biluoco formations. Q, quartz; F, feldspar; L, lithic fragments (Lm, metamorphic; Ls, sedimentary; Lv, volcanic). The black arrows highlight the compositional change consequent to the late Bathonian tectonic event.

Sandstone compositions are different in the north Biluoco and Suorika sections of the Abushan Formation. In the north Biluoco section, sandstones of the Abushan Formation are litho-quartzose and quartzo-lithic volcaniclastic (average composition QFL = 46:9:45, LmLsLv = 3:24:73) (Table S2). Intermediate to felsic volcanic rock fragments prevail over limestone and minor sandstone clasts. In the Suorika area, sandstone samples are litho-quartzose and quartzo-lithic sedimentaclastic (average composition QFL = 51:47:2, LmLsLv = 8:89:2) (Table S2), with dominant limestone and only minor sandstone and volcanic rock fragments.

Fig. 5. The image drawn with the merged line segments

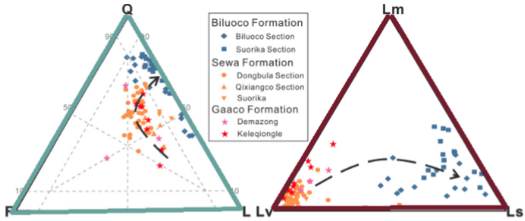

Figure 6. Ternary diagrams for sandstones of the Sewa, Gaaco, and Biluoco formations. Q, quartz; F, feldspar; L, lithic fragments (Lm, metamorphic; Ls, sedimentary; Lv, volcanic). The black arrows highlight the compositional change consequent to the late Bathonian tectonic event.

Sandstone compositions are different in the north Biluoco and Suorika sections of the Abushan Formation. In the north Biluoco section, sandstones of the Abushan Formation are litho-quartzose and quartzo-lithic volcaniclastic (average composition QFL = 46:9:45, LmLsLv = 3:24:73) (Table S2). Intermediate to felsic volcanic rock fragments prevail over limestone and minor sandstone clasts. In the Suorika area, sandstone samples are litho-quartzose and quartzo-lithic sedimentaclastic (average composition QFL = 51:47:2, LmLsLv = 8:89:2) (Table S2), with dominant limestone and only minor sandstone and volcanic rock fragments.

Fig. 6. The image drawn with the triangle boxes

experiment is checked manually. For the results of a triangle coordinate diagram, there are four possible kinds of results: correct, wrong, missing, and extra.

"Correct" means that the box given by the method can be positioned to the frame of the triangle coordinate diagram, and does not include other area that do not belong to the diagram; "Wrong" means that the number of triangle boxes given by the method is the same as the actual number of triangle coordinate diagram, but the given triangle boxes are where there is no triangle coordinate diagram or the given triangle box does not overlap with the triangle coordinate diagram; "Missing" means that the number of triangle boxes given by the method is less than the actual number of triangle coordinate diagrams, that is, some triangle coordinate diagrams fail to be given the corresponding triangle boxes on the result page; "Extra" means that the number of triangle boxes given by the method exceeds the number of actually existing triangle coordinate diagrams on the page.

4.3 Experimental Results and Numerical Analysis

Through experiments and manual verification, there are 112 triangle coordinate diagrams in 43 pages. The method in this paper gives 76 diagrams position results, of which 75 are "correct", 37 are "missing", 1 is "extra", and no "wrong"; The precision rate of the method is 98.68%, and the recall rate is 66.96%.

Experimental results show that this method has high precision, but low recall. The higher precision is attributed to the strict decision rules of the method. However, because of this, the strict decision rules have a certain impact on the recall rate. For example, when the data points are close to the frame of triangle coordinate diagram, when detecting the line segment of the frame, because of the existence of the data points on the frame, the frame line segment may be recognized as a multi-segment line segment or not a line segment (in order to avoid this problem, we introduce the step of line segment merging. However, this cannot completely avoid the problem), which will have an impact on the recall rate.

5 Conclusion

This paper solves the problem of academic literature triangle coordinate diagram localization. A new triangle coordinate diagram localization method is established by line segment detection and line segment merging. The purpose of locating triangle coordinate diagram in academic literature is realized. This method regards the problem of diagram localization as an image processing problem: Compared with the method based on the coding information in PDF format literature, this method can solve the problem of literature diagram localization in the scanning format of image; Compared with the method based on deep learning, this method does not need a lot of manpower to label data and can save manpower cost.

In this method, after detecting the position of line segment by line segment detection, the line segments are merged. Due to the merging of line segments, the wrong detection of line segment group can be avoided.

Nevertheless, there are still some improvements needed for this method.

(1) In the line segment merging step, it is necessary to manually set an appropriate threshold as the basis for line segment merging or not. Clustering the length of line segments and setting the threshold based on it may be a better way to make the generalization ability of the method stronger.

(2) When judging the triangle coordinate diagram, the model adopts the rules set manually to filter the line segments. This means that when dealing with other types of triangle coordinate diagram, it needs to be set according to the specific conditions of those diagrams. For example, this method can not effectively identify asymmetric triangle coordinate diagrams, triangle coordinate diagrams with inclined bottom edges or inverted triangle coordinate diagrams, which is one of the problems to be solved in the follow-up.

Acknowledgement. This research is supported by the National Natural Science Foundation of China under grant no. 42050102.

References

1. Rong, H., et al.: Integrated framework and visual knowledgometrics exploration for analyzing visual resources in academic literature. J. China Soc. Sci. Techn. Inf. **36**(2), 141–151 (2017)
2. Hao, F., et al.: Research on reading effect of the information diagram in the data news: evidence from the eye movement. Librar. Inf. Serv. **63**(8), 74–86 (2019)
3. Lee, P.S., West, J.D., Howe, B.: Viziometrics: analyzing visual information in the scientific literature. IEEE Trans. Big Data **4**(1), 117–129 (2016)
4. Apostolova, E., You, D., Xue, Z., et al.: Image retrieval from scientific publications: text and image content processing to separate multipanel figures. J. Am. Soc. Inform. Sci. Technol. **64**(5), 893–908 (2013)
5. Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.): ECCV 2016. LNCS, vol. 9909. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1
6. Ma, Y., Tung, A.K.H., Wang, W., Gao, X., Pan, Z., Chen, W.: ScatterNet: a deep subjective similarity model for visual analysis of scatterplots. IEEE Trans. Visual Comput. Graph. **26**(3), 1562–1576 (2020)
7. Yu, C., Levy, C.C., Saniee, I.: Convolutional neural networks for figure extraction in historical technical documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 789–795 (2017)
8. Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation of historical document images with convolutional autoencoders. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1011–1015. Tunis, Tunisia (2015)
9. Simon, A., Pret, J.-C., Johnson, A.P.: A fast algorithm for bottom-up document layout analysis. IEEE Trans. Pattern Anal. Mach. Intell. **19**(3), 273–277 (1997)
10. Clark, C., Divvala, S.: PDFFigures 2.0: Mining figures from research papers, 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), pp. 143–152. Newark, NJ, USA (2016)
11. Ma, Y.X., et al.: ScatterNet: a deep subjective similarity model for visual analysis of scatterplots. IEEE Trans. Visual Comput. Graph. **26**(3), 1562–1576 (2020)
12. Yu, C., Levy, C.C., Saniee, I.: Convolutional neural networks for figure extraction in historical technical documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 789–795. Kyoto, Japan (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. Las Vegas, NV, USA (2016)
14. Li, P., Jiang, X., Shatkay, H.: Extracting figures and captions from scientific publications. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1595–1598. Torino, Italy (2018)
15. Gioi, R.G., et al.: LSD: a fast line segment detector with a false detection control. IEEE Trans. Pattern. Anal. Mach. Intell. **32**(4), 722–732 (2010)