International Journal of Knowledge and Language Processing www.ijklp.org
KLP International ©2015 ISSN 2191-2734
Volume 6, Number 3, **2015** pp.48–62

Review of Research on Table Extraction in Scientific and Technical Literature

Lijun Sun and Yao Liu
Institute of Scientific and Technical Information of China
No. 15, Fuxing Road
Beijing 100038, P. R. China
sunlj2014@istic.ac.cn

Received August 2015; accepted September 2015

ABSTRACT. *As a multi-modal data, scientific and technical literature resources are rich in a variety of modes of information in text, images, formulas, tables, audio, video, etc. These information complements each other explanation, which helps users to fully understand the knowledge of scientific and technical literature. This paper describes main research methods and key technologies of table recognition base on image, web and in PDF and concludes their detection, otherwise, also refers to related research in the field of multimodal field as well as existing applications of table recognition technologies. Finally, the paper presents a future research direction.*

Keywords: scientific and technical literature, multi-modal, table extraction, semantic relativity

1. **Introduction.** As a multi-modal data, scientific and technical literature resources are rich in a variety of modes of information in text, images, formulas, tables, audio, video, etc. These information complements each other explanation, which helps users to fully understand the knowledge of scientific and technical literature. Compared with the traditional dominance of text information, the tables of which, as an important form of expression, descripts the text and you want to express more simple and intuitive, and complements with text [1]. In the information technology and network of twenty-first century, a lot of information is shown in the form of electronic documents. Millions of paper documents are being produced everyday, adding a never ending wealth of information to the human society. Practical use of these documents demand indexing, viewing, printing and extracting the intended portions in a fast and flexible way through electronic media. With the maturity of the document image analysis such systems are coming in the market. These include digital document libraries, vectorization of

engineering drawings and form processing systems [1, to name a few.

With the extensive application of tables, workload of table processing is no longer decided only by artificial statistics, but more dependent on the computer. The automatic input, storage and management have become an important part of intelligent processing document, and the research on table extraction has become a hot spot. The table extraction technology involves a number of different areas, such as financial statements, tax returns, student achievement table, identification card management system, license plate recognition, paperless evaluation system, terminal device development, and so on. If the table extraction technology is mature, with a table extraction system which has a high accuracy and high efficiency, it can greatly accelerate the speed of information input, improve work efficiency, and produce the huge economic benefits. Extraction of table information helps us get a large number of intelligence data, and also plays an important role in information science. Single modal analysis semantic understanding of information and comprehensive multi-modal information generated may be biased, thus using a variety of modes of complementary information on the form semantic understanding to get table information, and use multi-modal information to complement each other to improve search accuracy rate. Scientific and technical literature exists in various forms， maybe in handwritten documents, printed documents, even in web ,and any possible formats, so it is critical to learn table extraction technologies in every format of document.

2. **Related Research.** The research and development of table recognition technology is latter than the document recognition technology, Because of the lack of domestic high-speed scanning agents and few people knowing, table recognition market demand has not yet mature, and most of the people put effort in the printed text recognition technology. Along with acknowledge of the agencies for table recognition further understanding and mature OCR technology, coupled with the increase in market demand, people put more time and energy on research of form recognition technology. The present study is not mature for different formats, although there have been tireless efforts of the research and enough progress than before, we have not found a good common form recognition. With the social development, forms will be used in all walks of life, a lot of information need automate processing, thus the study on table recognition technology is an inevitable trend.

2.1. **Research on Table Extraction based on Image Processing.** Currently the main way to preserve most of the resource information is the document information, as the most convenient and consistent with the human habit of paper documents with many other documents irreplaceable advantages. Paper documents will continue to occupy an important position in future life [2] [3]. There are a surprising number of paper documents to be processed every day, manpower alone has been unable to meet people's needs, so it is an inevitable trend of making use of computer to do digital processing, storage and management for a variety of document information. At present, the main method for the identification of paper documents is using OCR technology, which obtains the paper text image information by scanning and imaging and other optical input, using a variety of pattern recognition algorithms morphological to analyze the text morphological characteristics, judge the standard encoding of Chinese characters, and storage according to a common format in a text file. It is based on image processing for the image document, and initial studies on table recognition are based on the image too, the main steps are as follows: image pre-processing, spreadsheet line extraction and consolidation, character extraction, character OCR processing (Figure 1) [4]. Now research on table extraction based on image processing is more mature, people more care about key technologies being improved and perfected.

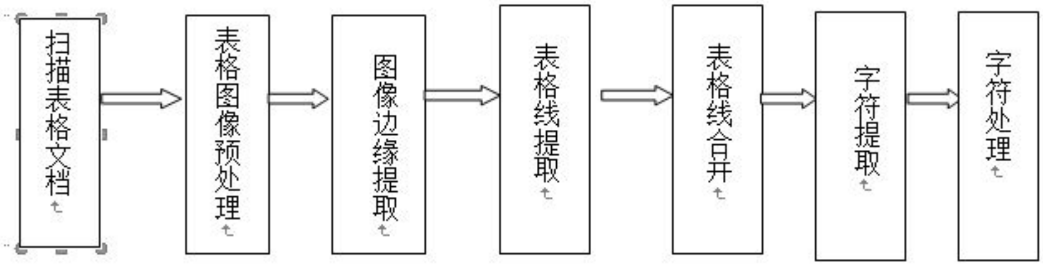

FIGURE 1. MAIN STEPS

For example, existing common image binarization algorithm: clustering method of Ridler and Calvard, local adaptive thresholding method of Sauvola and Pietaksinen, threshold segmentation method based on the histogram valley point, Otsu method (Ostu algorithm), the maximum entropy method, Hwang and Fu-based multi-scale canny edge binarization method [5-9]. However, these methods are relatively simple, in practice it is unable to better adapt to various situations in different conditions, so there are still scholars continuing to study to improve binarization processing technology. Document [10] proposed a nonlinear contrast enhancement and LoG Operator mixed binarization method (referred Hybrid method).

There are many tilt correction methods now, and they can be summarized into five main categories: Hough transform, lateral horizontal projection method, the Fourier transform method, k- nearest neighbor clustering method, and the linear fitting method [11] [12]. Furthermore, the former four methods are more time consuming, and in addition to Hough transform method, other methods are not guaranteed to achieve better accuracy for able document image which containing handwritten digits. But the tilt correction image is the key technology of image preprocessing, directly affecting the character extraction, thus people pay more attention on such research. Document [13] proposed a tilt angle detection algorithm which based on connected region's minimum bounding matrix. Document [14] proposed the method is to tilt detection for all in image, then take the largest number of straight direction as the tilt direction of the entire image. Document [15] made detailed introduction for several tilt angle detection methods, and compared them through experiment.

Table line extraction is one of the key technologies of the image document recognition， some detection methods proposed for the table, such as a method based on a single communication link, extracting form lines based on a search, a method based on a block adjacent graph, etc. [16-18] And some for words lines overlap extraction, which can be roughly divided into two categories: one is to remove the grid lines, and according to the properties of the overlap region to recover lost stroke. The other is to classify the possible overlap of words and lines in advance, and give each class a treatment, detect a match which based to a separation. For example, paper [19] refers to lots of Korean table analyzed, summarizes 13 kinds of Korean characters and form lines overlapping type, and matches the detected overlapping with one type, then processing. Paper [20] and paper [21] also belong to this type of algorithm, the former divided intersect angle of strokes and lines into three: Large angle, small angle and almost tangentially, the latter based on an overlapping manner between strokes and lines (separation and intersection) dividing the pixel area on the line into protected area and erased area, then processed separately.

It is more difficult to accurately locate and detect table character for table extraction, there are two main approaches: one is the use of a priori knowledge, and the other is to locate through the table line detection. Paper [22] proposes a form character location and extraction method based on COI extraction and partial classification recovery of fracture stroke.

2.2. **Research on Table Extraction based on Web.** Traditional table information extraction research looks at the ASCII file or forms obtained by OCR. In the late 1990s, with the expansion of web information, and gradually the web form information extraction task has been raised. The main form of web pages in the table includes HTML, PDF, image, TXT, XML, etc., currently web form information extraction research focusing mainly for HTML table.

It is relatively easy to identify this type of form, because the form of rows and columns of cells are tagged by the HTML tags. When extract such types of tables, we can get the desired results according to the table structure information that is "attribute - value" to expand the extraction. For example, the first official study of the extraction process web forms includes forms positioning, table structure recognition and "attribute-value" pair of extraction [23]. Document [24] further distinguished true table according tables detected by the label. Another reference [25] using the open source Mozilla browser box model rather than HTML tags to detect form, box model stores all the pages rectangle enclosing node information, then XY cutting method is used to detect the table area. Paper [26] refers to identify forms by analyzing web pages CSS2 visual rectangle.

2.3. **Research on Table Extraction in PDF.** Fixed layout of the format document is used for written documents, such as official publications, publishing documents, archives and the like. A lot of scientific and technology literature is present in the form of PDF, which is based on visual and does not exist in table format, the user can visually see the table generally results from the show, but was unable to obtain information directly from the document table format. PDF rendering the decoded text information "Text flow" feature, there are no borders stream node information, symbol or string information only, and no clear logic, it is difficult to apply past experience to identify the form of a PDF document.

Because of different author′s choreography, PDF format tables in scientific and technology literature may exist in the form of images, or in the form of text.

Early table recognition of layout document is to convert documents into image format, and then use a method based on image processing to carry out, until after 2005, it began to appear on the layout documents table recognition research. There are two main methods for PDF document identified, one approach is the use of tools to convert PDF documents into other formats, and then to form recognition and information extraction, another method is to directly table recognition in the PDF document.

Early study of PDF table recognition is to convert documents into image formats, such as literature [27] proposed a method that extract PDF potential table in 2002，specific steps are for post-PDF conversion image. The advantage of this algorithm is to convert PDF to TIFF image, so that the table is converted to an image-based recognition, you can use OCR technology. At the same time the problems are obvious, the original PDF information may be lost after transformation, and the recognition rates an image transformed have yet to be studied. Paper [28] also refers to convert PDF table into image document, then makes use of existing research results, this method will lose a wealth of original information. Document[29] proposed a table in the PDF structured approach, first convert PDF to HTML or XML format by pdftohtml tool， then identify and resolve the table, this usually lose the original information.

Often faced with the problem of losing the original data when we change the PDF into other formats, so researchers tend to extract tables directly in the PDF. Document [30] proposed a PDF table metadata extraction algorithm, the paper refers to a location analysis and keyword matching based methods to determine the table cell contents, identify table structure and detect table signs. The advantage of this method is that table recognition without the aid of a table image line, but the disadvantages are also obvious, algorithms are too theoretical, and there is no mention of how to identify the start and end positions of the table. Otherwise, it did not give data structure of table entity. Document [31] proposed an algorithm of detecting the position of the PDF by the image line. Detection of PDF document page "sparse line" to determine the content of the headings, tables, footnotes and so on with the "sparse line" feature of the text layout information. This method is to find a table the "sparse line" in detect position more focused and accurately. Paper [32] draws a table extraction based on image recognition algorithm of thinking, combined with the current PDF document table recognition and reproducible method, it proposes a new method for identifying and reproducible table. Separated by parsing the PDF document to text stream, and then to raster processing for text flow nodes which were extracted and marqueed, and finally return to the table of contents for the sequence output in HTML format. This approach lets zoning process and stored procedures separation, so that the identification process is more clear, reducing the complexity of the algorithm, but the string is not enough to identify integrity problems that will arise.

2.4. **Multi-Modal Semantic Relativity.** As a multi-modal data, scientific and technical literature resources are rich in a variety of modes of information in text, images, formulas, tables, audio, video, etc. These information complements each other explanation, which helps users to fully understand the knowledge of scientific and technical literature.

The concept of multi-modal is relative to the single-mode (unimodal or single-modality), Multi-modal studies generally refers to the use of different modes of information between two or more to solve a specific issue, there have not been a clearly generalized deny, multi-modality medical diagram Image registration refers to obtain two or more images from the different sensors, different time, different angles, then achieve best match processing. First appeared in the literature on multi-modal is study on pattern recognition of multi-modal test in 1968, after that in 1970 a paper about study on multi-signal detection function, proposed the concept of dual signal (bimodal signals), With respect to a single signal (unimodal signals). At the same period, multimodal also appeared in medical multi-modal therapy, biological systems and so on multi-modal learning and other fields. In the late 1990s, research on multi-modal gradually increased, applications are also more widely.

Multi-modal is similar with the concept of "multi-media", which refers to a human-machine interactive information exchange and media that is a combination of two or more medias, and multi-media include continuous media data (such as video, audio, etc.) and discrete media data (such as text, graphics etc.). It is not simply a composite of various media, but a combination of the text, graphics, images, moving videos and sound and other forms of information, and comprehensive treatment and control by computer, can support to complete a series of interactive operations. Currently, multimedia technology in the scientific data retrieval and disposal, management, business applications, education and vocational training, entertainment and other fields have a wide range of applications, which are mostly surrounded study about how to provide users with a better multimedia information service, which is a broad audio-visual services to expand. For example, multimedia learning is the focus of educational psychology, educational technology research and cognitive psychology, and according to human learning characteristics, combining the advantages of multimedia, organizing and arranging multimedia presentations effectively, promoting learners cognitive processing in the condition of multimedia. The multi-modal study focuses on the same goal by the different features or the same feature in different information comparisons and integration to solve a specific problem, at present mainly used in medical image registration and fusion with, biometric identification, images, audio, video, office management and retrieval, discourse analysis and other fields.

Multimodal is a relatively new field of study, because of the emphasis is to use different information of modes to problem-solving approach, not limited to a particular subject area, o the current domestic and international research involves a number of different modal factors, such as images, video, audio, biometric, speech expression, etc., there are variety of research areas, including the following aspects: multi-modality image automatic indexing and retrieval, multimodal medical image registration and fusion, multi-modal identification, classification and retrieval of multi-modal video, etc.

2.4.1. **Multi-Modality Image Automatic Indexing and Retrieval.** Multi-modality image automatic indexing and retrieval have got more and more attention in recent years. Image for people to understand the information has an important complementary role, with the accumulation of amounts of the picture in the Internet, what we can or not effectively organize the vast amounts of data and accurately search the desired image becomes an important research topic. Making use of multi-modal to complement information each other to improve search accuracy rate is the trends of the existing image search engines potential. Research on image semantic features includes automatic image annotation and image retrieval two aspects. Encountered with big data, multi-modality characteristics of internet image search, image-search-image, text-search-image and text image search image systems are all unsatisfying. In order to retrieve as many and as comprehensive images as possible, it's necessary to study on multi-modality fusion in internet image search. Internet mage search is a kind of multi-modality learning problem in essence. Many algorithms and ideas on them have emerged vector quantization or co-occurrence model, machine translation model, relevance model, structure model with class information, multi-label learning, complementary multi-modality fusion, multi-modality fusion based on matrix factorization, multi-modality fusion based on harmoniums, multi-modality fusion based on alignment learning, multi-modality joint learning, multi-modality learning on agreement, and multi-modality learning driven by big data[33].There are many ways we can learn.

2.4.2. **Multimodal medical image registration and fusion.** Multimodal medical image registration and fusion, is a hot issue in biomedical engineering. With the development of medical imaging and computer technology, the position of medical image in applications is increasingly heavy. However, what from a single image can not gain comprehensive diagnostic information, human space vision will shadow the accuracy of the results, so the study on multimodal medical image registration and fusion has been widespread concerned. Multimodal medical image registration and fusion mainly concentrated on the same patient at different times, under different sensors or two or more images were obtained under different conditions to registration and fusion, of whose method and key technologies to advance, related to the digital image processing, computer graphics and knowledge in the medical field, it is important application for computer graphics and image processing in biological project areas.

2.4.3. **Classification and retrieval of multi-modal video.** Classification and retrieval of multi-modal video, is mainly used in video event detection, classification and retrieval. With the rapid development of multimedia technology and internet application, multimedia data especially that of video data explodes. Requirements of efficient video retrieval become more and more important. Efficient technology of video retrieval can greatly help people obtain entertainment on internet and promote the quality of lives.

Nowadays, textual retrieval performs well. People can retrieval some relevant textual contents on internet by using baidu or google. Compared with the textual data, the structure of video data is more complicated: scene groups, scene, shot, frame. There are many kinds of feature information contained in video data, e.g. text, image, sound, which makes the process of the video more difficult and doing efficient video retrieval becomes the challenge. Actually people can retrieval the relevant video data by means of the features contained in video data. Many methods of video retrieval have been proposed up to now. In early days, people just did video retrieval by means of textual or image information singly. The textual based retrieval can promise high value of recall, while the image based retrieval performance well when the query topic has something to do with visual scene. Generally, the single-feature-used video retrieval cannot perform well, so people take all kinds of features into account. Different feature information has different advantages in video retrieval, making use of the technology of machine learning could promote the performance of video retrieval. According to different features, we can make several sub-retrieval modules. Nowadays many researches focus on how to apply suitable technology of machine learning on fusing and training these sub-retrieval modules. Different methods of machine learning have been used in video retrieval, rather their performances seemed not very well. The main reason of the unacceptable performance is that we have not fully extracted the information contained in video data. If we only make efforts on machine learning and neglect the video data itself, the performance of video retrieval cannot be promoted.

2.4.4. **Formula Recognition and Retrieval.** Formula recognition is another hot point, related research gradually increased in the 1990s. There are two mainly two: mathematical formulas recognition based on image and mathematical formula retrieval faced network. Distinguishing formula in the scanned image document is a mathematical formula to identify research hotspots. Since the scanning text file image includes not only a mathematical formula, also contains ordinary text, graphics and other objects, so first of all is to determine the position of the image, which based to cut and identify the symbols of formula, Then get the layout structure of formulas and semantic meaning through a variety of analytical methods. Finally, the results analyzed need to be output in accordance with a format (such as LaTeX), to achieve the purpose of the recognition result reuse. Depending on the characteristics of the category, mathematical formulas recognition based on image can be divided into three: character-based method, layout-based method and image –based method. Mathematical formulas used in the network environment is more and more widespread, People need to get the relevant information by searching for mathematical formulas. However, due to the special structure of mathematical formulas, you can not use the standard natural language processing method to process mathematical formulas, so mathematical formula retrieving is difficult. At present, for mathematical formulas retrieval there are two ways, one is the string representation of formula generation first, then by conventional information retrieval method for processing. The other is a structure-based method. But the two approaches still have its defection.

3. **Related Technology.**

3.1. **Optical Character Recognition.** OCR is the abbreviation for Optical Character

Recognition， which is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text. It is widely used as a form of data entry from printed paper data records, or any suitable documentation. It is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

OCR technology is mainly used for the identification of paper documents, the concept "Optical Character Recognition" was first put forward by German scientist Tausheck in 1929, then in the 1960s and 1970s, people began to study OCR all over the world. Around 1960 began to study the basic OCR recognition theory, which aimed at digital object initially, and there were some simple products until 1965-1970, such as ZIP code identification system, which identify messages on zip code printed to help the post offices work as a regional division letter role. Chinese study on OCR technology started later, and began numbers, letters and symbols recognition studies in the 1970s, even in the late 1970s began to study Chinese character recognition. In domestic, the earliest to do OCR technology research and products are TH-OCR and Hanvon.

OCR engines have been developed into many kinds of object-oriented OCR applications, such as receipt OCR, invoice OCR, check OCR, legal billing document OCR. They can be used for data entry for business documents like check, passport, invoice, bank statement and receipt, a lot of text information, automatic number plate recognition, electronic scoring system, automatic insurance documents key information extraction, extracting business card information into a contact list, make electronic images of printed documents searchable, and even for blind and visually impaired users as an assistive technology. OCR technology is an important milestone, which greatly improving the efficiency of data input, storage, retrieval and processing, and promoting the study of table recognition technology.

3.2. **Electronic document classification.** Currently the classification of table documents is not clearly defined, not only can be classified in accordance with areas the table contents involved, but also according to the layout, also can be divided according to research disciplines and areas covered. For example, according to the organizational form of data, they can be divided into binary files and ASCII documents, however, considering the use of them, they would also be divided into system documentations, user documentations and library documentations.

Considering the format structure and parsing angle, the electronic documents can be divided into layout documents and flow documents. The features of layout documents are fixed layout, a stable version, and in the process of using electronic documents showing the effects that do not change duing to hardware and software environment, the change of operator, finally in the format, layout, font, size and other aspects remaining fully consistent with the paper document. In other sides, the features of layout document format make it an ideal document format for electronic document, the digital information dissemination and archiving. At present, the layout of the electronic document has XPS, PDF, and CEB. Flow documents are mainly stored in the logical data, the characteristics of the document is typeset, adapting to the current display environment, presenting different layout effects according to different environmental impacts. Typical flow document has Microsoft ‘s launched DOC format, such electronic document table generally has an independent structure, which leads to easy conversion, edition and reuse. Therefore, such documents table recognition has rarely studied.

According to the document object type, documents can be divided into: the image documents, web documents, plain text documents, layout documents. Where most researches on which are image documents and layout documents.

3.3. **Research Content of Table Recognition.** In general, the table identification involves three subtasks: table positioning, structural analysis, semantic understanding. Table positioning, the aim of which is to detect the presence of a table object or not for the given document page, and determine the boundaries of the table area. Table positioning is the first step in a document form recognition process and the most critical step, the positioning accuracy will directly affect the subsequent table structure analysis, understanding and other steps of recognition performance, so that researchers pay much attention on the document analysis and understanding in nearly two decades.

Existing research methods on table positioning are mainly divided into three categories: a method based on a predefined template, a rule-based approach (also called heuristics) and a statistical learning-based approach. The predefined template method refers to a set of form templates by the user or pre-defined, which according to their characteristics (such as the header, line description, empty, etc.), to identify new forms with similar characteristics; but provided that the form template with to be identified table having the same structure. Overall, deficiencies of such methods are more limited, you may need to add a new template at any time, so it generally applies only to a certain class form of a specific application domain. Rule-based approach, can be divided into two: according to form lines characteristic to summarize rules and according to the contents of the cell text layout features to inductive rules. The statistical learning-based approach there are: probability model, naive Bayes, Maximum Entropy, decision trees, support vector machines, CRFs, hidden Markov model, artificial neural networks. Although statistical learning based methods is more universal than the rule-based approach, inevitably it takes a lot of manually labeled training set, and feature selection, training and test sets building, classifier selection will cause different effects on the results of experiments, we need for more comprehensive summary.

Table structure analysis, including analysis of the physical structure and the logical structure at two levels. The former aims to determine the spatial location information of table rows, columns, and cells, the latter aims to analyze how the physical components of table interact to form a table. The existing table structure analysis studies mostly focused on the physical structure analysis, which based on terms spatial layout information in table, proposed rule-based method to divide rows and columns of table contents, and targeted solved across the ranks of cell division and other difficult issues. However, there are less research on logical structure analysis.

There is less study on table semantic understanding at present, as one of the future research directions, semantic understanding helps to recycle and store information, but also provides support for the construction of retrieval system. Table recognition is not limited to the table to identify itself, but should be combined with the semantic correlation between different modes, to build a structured semantic description system to achieve accurate identification form semantics.

3.4. **Application.** The research on recognition technology have promoted the development of digital information, accordingly, it generated a lot of applications to facilitate people's lives. The earliest domestic companies to do OCR technology research and development and formation of products are Tsinghua Unisplendour Wen Tong, Hanvon Technology, etc., and gradually formed HW, Wen Tong two-phase competition situation. Currently, form recognition technology has been applied to some OCR mobile terminals, such as document recognition, bank card identification, business card recognition, license plate recognition and identification terminal. Document OCR recognition also has data entry facility, digitizing systems. There are a variety of development and application of mobile reading devices, etc., even marking system. Form recognition technology has penetrated into every aspect of our lives. Domestic photovoltaic Engineering, Chongqing University, developed AV-100 automatic reader form, funding by the National High Technology Research and Development “863” program, got large-scale application in the agricultural census, and achieved significant economic and social benefit. Pattern Recognition and Intelligent Systems Laboratory, Beijing University of Posts and Telecommunications designed bank notes image processing and recognition engine in 2001, the project also received the support of the “863” program. There are still a considerable number of enterprises engaged in the development of products in this area in China, such as Beijing hanvon technology co., LTD. Its products related to automatic data entry function, such as Wang bank notes and business cards through the automatic identification system and other products. The commercial software Neusoft Smart Forms data entry and automated file management system that research and development by Shenyang Neusoft Co. Ltd, not only can realize intelligent identification tabular data, but also capable of automatic entry and file management data. Document analysis Recognition Technology that research and development by New York State University at Buffalo has achieved remarkable results. The main direction of the center is: Handwriting recognition, automatic handling envelopes and letters, intelligent character recognition, form processing and the like in Japanese document recognition. Pattern Recognition and Machine Intelligence Research Center, University of Canada formally established in September 1988. The center in character recognition, document image analysis and understanding as well as other aspects of natural language understanding of the world leading level. In addition, there are many foreign company engaged in product research and development in this field, the documentations, forms and the electronic documents that collected by Kofax company were transformed into accurate and recoverability of information. Seresoft company freed the people from the heavy manual labor, through superior data recording technology, and optimize the business process. Captive company provides the data inputs such as computer processing solutions, and optimize the information processing. Captive companies save time and costs while improving the accuracy of data acquisition.

4. **Key Technologies of Table Recognition.**

4.1. **Multimodal Information Resources Semantic Correlation.** As a multi-modal data, scientific and technical literature resources are rich in a variety of modes of information in text, images, formulas, tables, audio, video, etc. Study on different modes of semantic features and semantic relationships, table semantic analysis in context, finally achieve multi-modal integration of semantic features, which will help semantic understanding and tabular information extraction.

4.2. **Table Structure Analysis.** Table structure analysis, including analysis of the physical structure and the logical structure. The former aims to determine the spatial location information of table rows, columns, and cells, the latter aims to analyze how the physical components of the table interact to form a table. The current main task is to analyze the logical structure of the table.

4.3. **Table Semantic Understanding.** There is less study on table semantic understanding at present, as one of the future research directions, semantic understanding helps to recycle and store information, but also provides support for the construction of retrieval system. Table recognition is not limited to the table to identify itself, but should be combined with the semantic correlation between different modes, to build a structured semantic description system to achieve accurate identification form semantics.

4.4. **Table Retrieval.** Using complementary of multi-modal information to improve search accuracy rate is currently a research direction. Initially resolving table to obtain the initial semantic resources, according to retrieve to get some semantic resources which combined with network semantic resources, then get more resources to build a relatively complete semantic resources, and sort results, complete index, to complete the construction of the prototype system.

5. **Summary and Outlook.** As a multi-modal data, scientific and technical literature resources are rich in a variety of modes of information in text, images, formulas, tables, audio, video, etc. These information complements each other explanation, which helps users to fully understand the knowledge of scientific and technical literature. This paper describes main research methods and key technologies of table recognition base on image, web and in PDF and concludes their detection, otherwise, also refers to related research in the field of multimodal field as well as existing applications of table recognition technologies.

Multimodal is a relatively new field of study, involving a number of different modes of factors, and is not limited to a particular subject area, mainly including multi-modal automatic image indexing and retrieval, multi-modal medical image registration and fusion, multi-modal identification, classification and retrieval of multi-modal video information, multi-modal human-computer interaction systems research, multi-modal discourse analysis, multi-modal emotion recognition and so on. Learning multimodal related field of study, will help to understand the content and methods of multimodal research, and play an important role for multi-modal feature fusion method research.

Scientific and technical literature exists in various forms， maybe in handwritten documents, printed documents, even in web ,and any possible formats, so it is critical to learn table extraction technologies in every format of document. Currently the main way to preserve most of the resource information is the document information, as the most convenient and consistent with the human habit of paper documents with many other documents irreplaceable advantages. Paper documents will continue to occupy an important position in future life. The main method for the identification of paper documents is using OCR technology, which obtains the paper text image information by scanning and imaging and other optical input, then storage according to a common format in a text file. Main steps are: image pre-processing, spreadsheet line extraction and consolidation, character extraction, character OCR processing. The form of web pages in the table includes HTML, PDF, image, TXT, XML, etc., mainly HTML, it is relatively easy to identify this type of form, because the form of rows and columns of cells are tagged by the HTML tags. However, in PDF, it will be more difficult. There are two main methods for PDF document identified, one approach is the use of tools to convert PDF documents into other formats, and then to form recognition and information extraction, another method is to directly table recognition in the PDF document. The technologies of table recognition in PDF have not been mature yet, there are amounts of work to do.

OCR is important in table recognition, it is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. So we have to pay attention to OCR technology. There are many applications of OCR, too.

The future work mainly focuses on table extraction in PDF, we can use some better technologies for image document or web document, our research contents contain: multimodal information resources semantic correlation, table structure analysis, table semantic understanding, table retrieval. Full advantage of multi-modal information can improve table recognition results, but also the inevitable trend of the development of table extraction.

Acknowledgement. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2013BAH21B02, and the Key Work Project of Institute of Scientific and Technical Information of China (ISTIC) No. ZD2015-19; the authors also gratefully acknowledge the helpful comments and suggestions of the reviewers,

which have improved the presentation.

REFERENCES

[1] Ruijia Wang, Yao Liu. Study on the Feature Extraction and Expression System of Multi-Modal Semantic Information for Scientific and Technical Literature [J]. Journal of Academic Libraries，2012, 30（5）：71-76.

[2] Yu Liu. Research on Printed Forms Recognition [D]. Harbin Engineering University，2014.

[3] Bofeng Yu. Recognition of forms and printed Chinese characters in the document [D].Harbin Engineering University, 2011.

[4] Ming Si. Research on form recognition [D]. Xi`an University of Science and Technology, 2009.

[5] T.W.Ridler, S.calvard. Picture thresholding using an iterative threshold selection method. IEEE 7rralls，SMC-8, pp.630-632, 1978．

[6] J.Sauvola and M.Pietaksinen. Adaptive document image binarization， Pattern Recogn． 33，225-236(2000)． In JOURNAL OF APPLIED RESEARCH AND TECHNOLOGY, 2012 (10),703-712.

[7] P. Moallem, N. Razmjooy. Optimal Threshold Computing in Automatic Image Thresholding using Adaptive Particle Swarm Optimization .In JOURNAL OF APPLIED RESEARCH AND TECHNOLOGY, 2012 (10).

[8] Otsu ’ s thresholding, image segmentation, picture thresholding, multilevel thresholding, recursive algorithm. ，Image Processing IEEE Transactions 0n Vol.10，Issue 8，Aug．2001:1152-1161.

[9] Hwang.Wen L and Fu.chang. Character extraction from documents using wavelet maxima, Image and Vision Computing. vol．16,Issue：5，April 27，1998，307-315.

[10] Xie, Liang. Research on Pre-processing and Character Extraction of Form Document Recognition [D]. GUANGZHOU：Sun Yat-sen University，2005.

[11] Yunhua Li, Huichuan Duan. The Form Withdraw and Slant Correction in Picture Files Based on Hough Transform [J]. INFORMATION TECHNOLOGY & INFORMATIZATION, 2007（6）：49-51.

[12] Hiroshi Shinjo，Eiichi Hadano，Katsumi Marukawa，Yoshihiro shima，and Hiroshi sako．A recursive analysis for form cell recognition．2001.

[13] R．Safabakhsh，S．Khadivi，“Document skew detection using minimum—area bounding rectangle” 2000，IEEE International Conference on Information Technology，2000，Pages：253-258.

[14] S. W. Lam，Javanbakht，S. N. Srihari, Anatomy of a Form Reader. Proc. Second hat’l Conf. Document Anal．Reeog，PP．506-509，Tsukuba，Japan，1993．

[15] Hwang.Wen L and Fu.chang. Character extraction from documents using wavelet maxima, Image and Vision Computing.vol．16,Issue：5，April 27，1998，pp．307_315.

[16] Yefeng Zheng, Changsong Liu, Xiaoqing Ding, Shiyan Pan. A Form Frame-Line Detection Algorithm Based on Directional Single-Connected Chain [J]. JOURNAL OF SOFTWARE2003，23(4)：790-796．

[17] Yefeng Zheng，Huiping Li，David Doermann．A Parallel-Line Detection Algorithm Based on HMM Decoding. IEEE，Transactions on Pattern Analysis and Machine Intelligence Volume 27，Issue 5，May 2005：777-792．

[18] Bin Yu, Anti K．Jain．A Generic System for Form Dropout．IEEE Transactions On Pattern Analysis and Machine Intelligence，1996,18(11):1127-1134.

[19] Yoo J Y．Kim M，Han S Y, Kwon Y B．Line Removal and Restoration of Handwritten Characters on the Form Documents． Proc of 4th International Conference． Document Analysis & Recognition，Ulm Germany,l997．

[20] Yefeng Zheng，Huiping Li，David Doermann． A Parallel-Line Detection Algorithm Based on HMM Decoding. Journal of Electronics and Information Technology, 2002，24(9)：1190-l196．

[21] Chongyang Zhang, Zhen Lou, Yong Xu. Detection and removal of form lines from bill images [J]. COMPUTER ENGINEERING AND DESIGN, 2005，26(7)：1778-1780．

[22] Xie, Liang. Research on Pre-processing and Character Extraction of Form Document Recognition [D]. GUANGZHOU：Sun Yat-sen University，2005.

[23] Chen H, Tsai S, Tsai J. Mining Tables from Large Scale HTML Texts[C]. Proceeding of the 18th International World Wide Web Conference, 2007:71-80.

[24] Yalin Wang, Jianying Hu. Detecting Tables in HTML Documents. In Proceeding of the 5th International Workshop on Document Analysis Systems，Princeton，NJ，2002．

[25] W.GaRerbauer B.Krupl，M.Herzog． Using visual cues for extraction of tabular data from arbitrary HTML documents． In Proceeding of the 14th International Conference on World Wide Web，pages 1000-1001，2005．

[26] Wolfgang Gatterbauer and Paul Bohunsky． Table extraction using spatial reasoning on the CSS2 visual box model． In Proceedings of the 21st National Conference on Artificial Intelligence(AAAI 2006)．AAAI，MIT Press，July 2006．

[27] H.C.Wasserman, K.YuKawa, B.K.Sy, K-L.Kwok, and I.T.Phillips. A theoretical foundation and a method for document table structure extraction and decompositon. In Document Analysis Systems, 2002, 291-294.

[28] Embley, DW; Lopresti, D; Nagy, G. Notes on Contemporary Table Recognition． 7th International Workshop on Document Analysis Systems, 2006.

[29] B.Yildiz, K.Kaiser, S.Miksch. pdf2table： A method to extract table information from PDF files.Proceedings of the 2nd Indian International Conference on Artificial Intelligence,2005:1-13.

[30] Ying Liu, Prasenjit Mitra, C.Lee Giles, Kun Bai. Automatic extraction of table metadata from digital documents.Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries,2006:339-340.

[31] Ying Liu, Prasenjit Mitra, C.Lee Giles. Identifying table boundaries in digital documents via sparse line detection.Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008:1311-1320.

[32] Jing Fang. Research on Table Detection and Structure Analysis from Fixed-layout Document [D]. Peking University，2013.

[33] Yanping Zhou. Multimodal Fusion in Internet image search [D]. University Of Science And Technology Of China, 2015.