Open access • Reference Book • DOI:10.1007/978-0-85729-859-1

# Handbook of Document Image Processing and Recognition — Source link

David Doermann, Karl Tombre

**Published on:** 21 May 2014

# Recognition of Tables and Forms

Bertrand Coüasnon, Aurélie Lemaitre

# Recognition of Tables and Forms

# 19

Bertrand Coüasnon and Aurélie Lemaitre

## Contents



**Abstract**

Tables and forms are a very common way to organize information in structured documents. Their recognition is fundamental for the recognition of the documents. Indeed, the physical organization of a table or a form gives a lot of information concerning the logical meaning of the content.

This chapter presents the different tasks that are related to the recognition of tables and forms and the associated well-known methods and remaining

B. Coüasnon (✉)
IRISA/INSA de Rennes, Rennes Cedex, France
e-mail: couasnon@irisa.fr

A. Lemaitre
IRISA/Université Rennes 2, Rennes Cedex, France
e-mail:couasnon@irisa.fr

D. Doermann, K. Tombre (eds.), *Handbook of Document Image Processing and Recognition*, DOI 10.1007/978-0-85729-859-1_20,

challenges. Three main tasks are pointed out: the detection of tables in heterogeneous documents; the classification of tables and forms, according to predefined models; and the recognition of table and form contents. The complexity of these three tasks is related to the kind of studied document: image-based document or digital-born documents. At last, this chapter will introduce some existing systems for table and form analysis.

**Keywords**
Detection • Digital-born documents • Form • Identification • Image-based documents • Structure detection • Table

## Introduction

### History and Importance

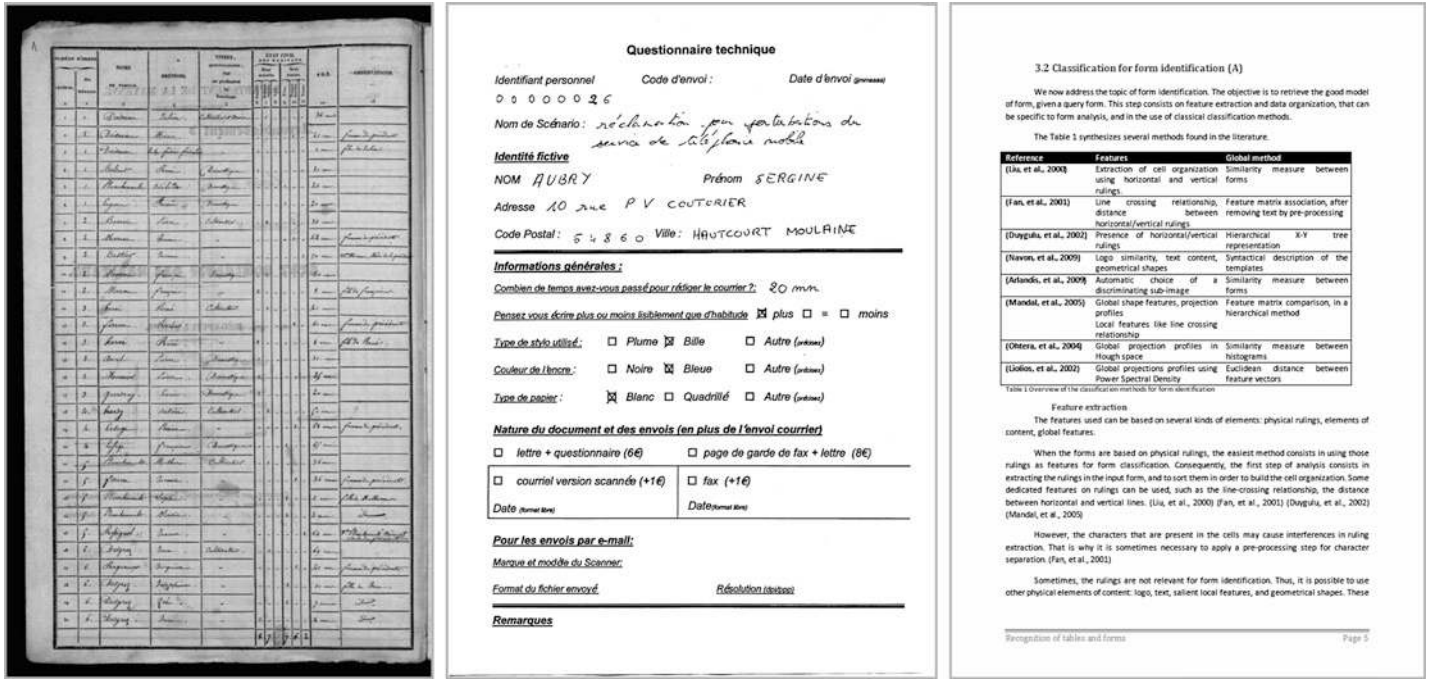
Tables and form are essential in the presentation of the information and the organization of contents in documents. Thus, they have been studied since the beginning of structured document analysis. For a long time, the research has been focused on the extraction of low-level geometric information from scanned images or paper tables. Most recent researches are focused on the analysis of tables and forms in electronic support, in order to obtain a higher level of table understanding.

The recognition of tables and forms is very important in document analysis. Thus, the physical organization of tables and forms provides some information on the logical organization of the content. Consequently, the result of a step of recognition of tables and forms can really improve the recognition of the handwriting.

Moreover, tables and forms are present in a large variety of documents (Fig. 19.1): administrative documents, invoices, question forms, old documents, etc. Consequently, a system that aims at recognizing open documents must deal with the recognition of tables and forms.

### What Is a Table?

The definition of a table has been discussed in many papers. Indeed it is quite easy to have an idea of what a table is, but a precise and formal definition is more difficult. Tables are the prevalent means of representing and communicating structured data. They may contain words, numbers, formulae, and even graphics [1]. They can contain printed information or handwritten information (mainly in archives documents). Tables are a 2D cell assembly where data type is determined by either horizontal or vertical index [2]. Costa e Silva et al. [3] propose a definition of a table as a "graphical grid-like representation of a matrix $M_{i,j}$ where: (1) each element i, j of the matrix is atomic; (2) there are linear visual clues, i.e the elements of each row i (column j) of the matrix tend to be horizontally (vertically) aligned;



Questionnaire technique

Identifiant personnel: 0 0 0 0 0 0 2 6 — Code d'envoi: — Date d'envoi (jjmmaaaa)

Nom de Scénario: réclamation pour factures de service de téléphone mobile

Identité fictive

NOM AUBRY — Prénom SERGINE

Adresse 10 rue P V COUTURIER

Code Postal: 5 4 8 6 0 — Ville: HAUTCOURT MOULAINE

Informations générales :

Combien de temps avez-vous passé pour rédiger le courrier ?: 20 mn

Pensez vous écrire plus ou moins lisiblement que d'habitude ☒ plus ☐ = ☐ moins

Type de stylo utilisé : ☐ Plume ☒ Bille ☐ Autre (préciser)

Couleur de l'encre : ☐ Noire ☒ Bleue ☐ Autre (préciser)

Type de papier : ☒ Blanc ☐ Quadrillé ☐ Autre (préciser)

Nature du document et des envois (en plus de l'envoi courrier)

☐ lettre + questionnaire (6€) ☐ page de garde de fax + lettre (8€)

| | |
|---|---|
| ☐ courriel version scannée (+1€) | ☐ fax (+1€) |
| Date (format libre) | Date (format libre) |

Pour les envois par e-mail:

Marque et modèle du Scanner:

Format du fichier envoyé — Résolution (dpi/ppp)

Remarques

3.2 Classification for form identification (A)

We now address the topic of form identification. The objective is to retrieve the good model of form, given a query form. This step consists on feature extraction and data organization, that can be specific to form analysis, and in the use of classical classification methods.

The Table 1 synthesizes several methods found in the literature.

| Reference | Features | Global method |
|---|---|---|
| (Liu, et al., 2000) | Extraction of cell organization using horizontal and vertical rulings | Similarity measure between forms |
| (Fan, et al., 2001) | Line crossing relationship, distance between horizontal/vertical rulings | Feature matrix association, after removing text by pre-processing |
| (Duygulu, et al., 2002) | Presence of horizontal/vertical rulings | Hierarchical X-Y tree representation |
| (Navon, et al., 2009) | Logo similarity, text content, geometrical shapes | Syntactical description of the templates |
| (Arlandis, et al., 2009) | Automatic choice of a discriminating sub-image | Similarity measure between forms |
| (Mandal, et al., 2005) | Global shape features, projection profiles<br>Local features like line crossing relationship | Feature matrix comparison, in a hierarchical method |
| (Ottera, et al., 2004) | Global projection profiles in Hough space | Similarity measure between histograms |
| (Lioios, et al., 2002) | Global projections profiles using Power Spectral Density | Euclidean distance between feature vectors |

Table 1 Overview of the classification methods for form identification

Feature extraction

The features used can be based on several kinds of elements: physical rulings, elements of content, global features.

When the forms are based on physical rulings, the easiest method consists in using those rulings as features for form classification. Consequently, the first step of analysis consists in extracting the rulings in the input form, and to sort them in order to build the cell organization. Some dedicated features on rulings can be used, such as the line-crossing relationship, the distance between horizontal and vertical lines. (Liu, et al., 2000) (Fan, et al., 2001) (Duygulu, et al., 2002) (Mandal, et al., 2005)

However, the characters that are present in the cells may cause interferences in ruling extraction. That is why it is sometimes necessary to apply a pre-processing step for character separation (Fan, et al., 2001)

Sometimes, the rulings are not relevant for form identification. Thus, it is possible to use other physical elements of content: logo, text, salient local features, and geometrical shapes. These

Recognition of tables and forms — Page 5

**Fig. 19.1** Various kinds of tables/forms: hierarchical table in archive document, mixed printed/handwritten form, and digital-born document containing a table

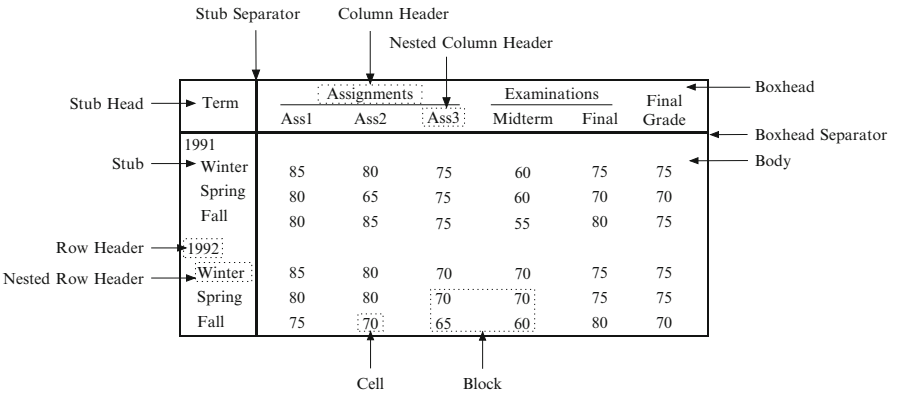



**Fig. 19.2** Terminology of parts of tables initially presented in Wang [4] and modified in [5]

(3) linear visual clues describe logical connections [...]; (4) eventual line art does not add meaning otherwise not present in the relative positioning of the cells in the table." This definition covers a large part of tables but has some limitations to include tables that can have recursive structure, with another table structure in a cell.

The terminology of the different parts of a table is presented in Fig. 19.2. A hierarchy of columns (resp. rows) can be define in the boxhead (resp. stub) with column (resp. row) headers and nested column (resp. row) headers. This hierarchy can be quite complex and not limited to two levels like in Fig. 19.2. Cells, columns, and rows can be delimited by physical rulings or by perceptive rulings built by alignments of cell contents.

In this chapter, tables are classified according to their complexity: simple 2D-tables and complex 2D-tables. Simple 2D-tables are limited to a matrix of cells with simple rows and columns with boxhead and stub limited to one level of hierarchy, like the table of Fig. 19.1 right. Complex 2D-tables can have some hierarchy of columns and/or rows, and may be recursive, like the table of Fig. 19.1 left. This separation, according to the table complexity, is important to see if a table recognition method is limited or not to simple tables.

## What Is a Form?

A form is a document with a predefined template mainly used for collecting information, with a one-to-one mapping between indices and data and no implication of regularity [2]. The data collected can be handwritten or machine printed, which produces segmentation difficulties with the preprint template. Even if forms are more often used for collecting data, it is important to notice that they can also be used to communicate this data.

When a form is made of rulings, it corresponds to a 1D-table, most of time horizontal, like the example of Fig. 19.3. Forms may contain 2D-tables. Ruled


**Fig. 19.3** Example of forms

versions of forms are usually a mix of 1D-tables and 2D-tables, with 1D-tables in a cell of a 2D-tables or a juxtaposition of 1D- and 2D-tables. When a form for collecting information is made of a 2D-table only, it can be recognized as a table.

## Table/Form Representation

Table and form representation is important for table-structured data extracted by document analysis. It depends of the precision of the recognition. Ordered lists of cells and rows (or columns) can represent the table matrix. For more complex tables or forms, graphs and trees can be used to integrate relationships between form elements [1].

For defining a representation, it is interesting to study the formal table model proposed by Wang [4] for generating table images from a table description. This model is the most complete table model in the literature [5]. It separates the logical part and the physical part. The logical part contains row and column hierarchy, while in the physical part, separator type, size, and content display (fonts, alignments, etc.) are specified for a cell or a set of cells.

## Objectives and Main Difficulties

The problem of table and form processing can be split into three themes: the detection of tables in documents that contain heterogeneous information, the classification of forms using predefined models, and the recognition of the tabular structure and its content. Table and form processing can be done in both image-based documents and digital-born documents.

### Detection

The problem of table detection/localization is fundamental in the analysis of heterogeneous document images. This task consists in finding *if* there is a table/form in a document and *where* the table/form in the document is. These two tasks are often joined. This topic concerns mixed text and table document images such as scientific papers or business and law books. But the problem is also addressed for more complex documents such as internal reports, financial statements, technical and commercial notes, magazines, scientific papers, tickets, and certificates.

The detection of table is as well sometimes necessary in the context of mixed printed/handwritten documents.

In the field of digital document processing (text, HTML, PDF), the table localization is a challenging task. Thus, in text documents, the tables are not always marked by physical rulings but by tabulations and spaces, which complicates their detection. Table detection can also be difficult in PDF documents. At last, the detection of tables in HTML documents requires a specific processing as the <TABLE> HTML tag is often used for both real meaningful tables and for the construction of the web page layout.

### Classification

Another task for table/form processing, which differs from the complete recognition of the document, is form classification. What is called form, in this context, are preprinted documents that are hand-filled. Thus, for the processing of the forms, the objective is to identify the type the candidate form belongs to and then to analyze the content of handwritten fields.

One may identify two levels of complexity for the classification. When the candidate form is exactly the same as the blank model, the difficulties are to deal with skew, shift, and the presence of handwriting. In more complex cases, the forms can vary inside of a given template: size of boxes, text, print quality, scales, etc. In these cases, the methods have to deal with the variation of templates and possibly with a reject option.

### Recognition

The most complex task in table processing consists in the recognition of the physical and the logical content. Recognition implies to understand the structural organization of tables to extract data and if necessary to reorganize this data.

The recognition of tables often requires specific systems that are adapted for the analysis of the tabular structures. Once the structure is detected, the recognition of the content can often be realized using the classical writing recognizers. However in complex documents, the recognition of the content is necessary to perform the structure recognition. Consequently the process of table recognition needs to combine both structure and content recognition.

## Table and Form Processing

The main content is organized as follows. Two kinds of documents will be studied: the image-based documents and the digital-born documents. For each kind of document, there will be three tasks: the detection, the classification, and the recognition.

## Table and Forms in Image-Based Documents

This section focuses on the analysis of images of documents. They are handwritten, printed, or mixed documents that have been digitized. There is no available electronic metadata for the recognition of these documents. For this kind of documents, three tasks are studied:
– The detection of tables/forms in heterogeneous documents
– The classification of forms into models
– The recognition of tables

### Detection

In document images, the simplest way for table detection and localization is to analyze structural features. It is possible to define rules or heuristics that are applied to detect the presence of tables in complex documents.

The easiest configuration for table/form localization is when physical rulings are present in the documents. Thus, the strategy for table localization consists in first extracting the rulings and then looking for a specific arrangement in lines, rows, and cells.

For the detection of rulings, the classical techniques of line segment detection can be used (see ▸Chap. 15 (Graphics Recognition Techniques)). Then, the extracted line-segments can be organized to match with a tabular structure. A tabular structure can be characterized by the presence of at least two parallel lines, having the same size.

Depending of the used technique, the ruling extraction may require a phase of preprocessing, like skew correction. Some methods are especially convenient for table analysis (▸Chap. 4 (Imaging Techniques in Document Analysis Processes) gives details on imaging techniques for document analysis). A recursive X-Y tree analysis of the page requires thresholds like the criteria to decide whether two parallel lines have the same length, or the size of a white strip to separate columns. These thresholds can be learned using the principle of a solving and optimization problem [6].

Some recursive structural approaches can also be used in regular documents. Thus, it is possible to realize a recursive decomposition of the document images into rectangles until they form regular cells. This approach is used by Ramel et al. [7] in the analysis of printed PDF documents.

In order to avoid specific learning, the morphological approaches enable to estimate the presence of vertical rulings, horizontal rulings, and the intersection of rulings [8] (Fig. 19.4). The morphological tools are particularly convenient for the analysis of hand-filled tables/forms. Indeed, the presence of handwriting complicates the detection of the rulings. The use of watershed transform, combined with statistical analysis, enables to deal with the noise due to the presence of text inside the cells [9] (Fig. 19.3).

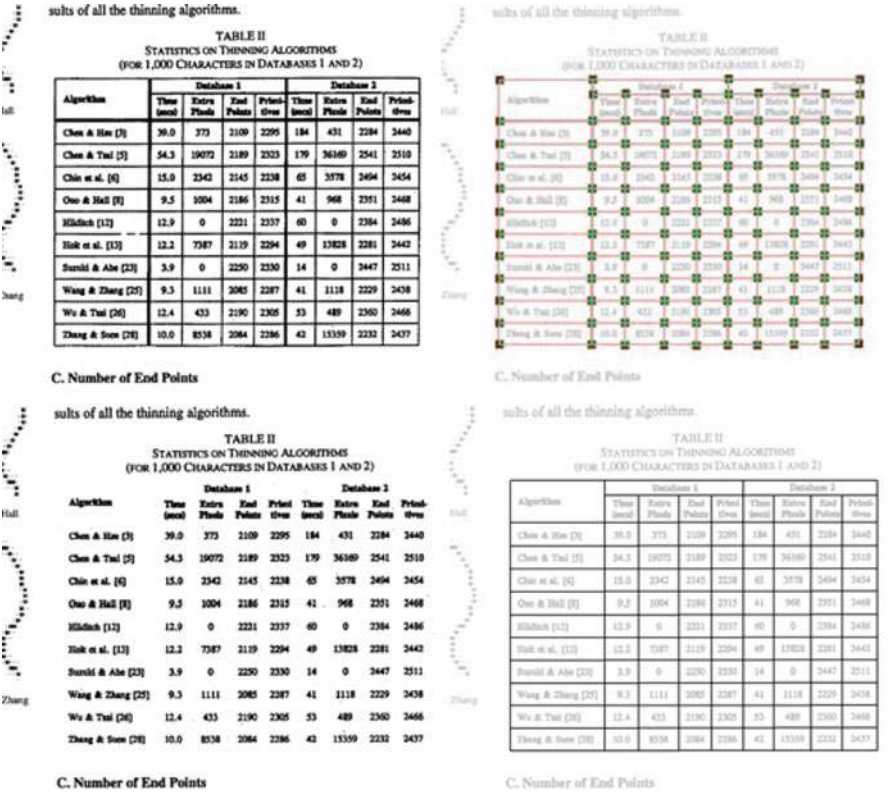



**Fig. 19.4** Table detection based on physical rulings (initial image, detected line intersections, image without rulings, final table reconstruction) [8]

In many cases, there are no physical rulings for the delimitation of each cell. Some rulings are sometimes present to delimit only certain columns or rows. Consequently, table detection must be based on other physical elements such as regularity of the positions, alignments, and presence of white spaces.

The simplest approach consists in the study of blocks that are horizontally or vertically aligned, with a certain threshold of similarity. The projection method can be used to build columns and rows. However, this method requires parameters and clean documents. This approach is followed by the T-Recs table recognition system [10].

Some other important information for table detection is the fact that the spacing between columns is regular. Thus, it is possible to group connected components into word and words into lines and then to segment the lines into columns depending on the interword space. This method does not need specific parameters [11]. However, it is dedicated for documents with Manhattan layout and may not work for complex documents with heterogeneous arrangements.

**Table 19.1** Example of forms

| Reference | Features | Global method |
|---|---|---|
| [12] | Extraction of cell organization using horizontal and vertical rulings | Similarity measure between forms |
| [13] | Line-crossing relationship, distance between horizontal/vertical rulings | Feature matrix association, after removing text by preprocessing |
| [14] | Presence of horizontal and vertical rulings | Hierarchical X-Y tree representation |
| [15] | Logo similarity, text content, geometrical shapes | Syntactical description of the templates |
| [16] | Automatic choice of a discriminating sub-image | Similarity measure between forms |
| [17] | Global shape features, projection profiles | Feature matrix comparison, in a hierarchical method |
| | Local features like line-crossing relationship | |
| [18] | Global projection profiles in Hough space | Similarity measure between histograms |
| [19] | Global projections profiles using Power Spectral Density | Euclidean distance between feature vectors |

As a conclusion, the structural methods based on rules and heuristics are convenient and widely used for table detection, more particularly when the table cells are delimited by physical rulings. However, in more complex documents, the analysis of the content could improve the table localization. Section "Detection" will show that some knowledge on metadata or the content of documents in digital-born documents enables to treat more complex documents.

## Classification for Form Identification

The objective of form identification is to retrieve the good model of form, given a query form. This step consists on feature extraction and data organization that can be specific to form analysis and in the use of classical classification methods. Some of features and classification methods are described in ▸Chap. 11 (Handprinted Character and Word Recognition).

Table 19.1 synthesizes several methods found in the literature.

**Feature Extraction** The features used can be based on several kinds of elements: physical rulings, elements of content, and global features.

When the forms are based on physical rulings, the easiest method consists in using those rulings as features for form classification. Consequently, the first step of analysis consists in extracting the rulings in the input form (e.g., with methods presented in ▸Chap. 15 (Graphics Recognition Techniques)) and to sort them in order to build the cell organization. Some dedicated features on rulings can be used, such as the line-crossing relationship and the distance between horizontal and vertical lines [12–14, 17].

However, the characters that are present in the cells may cause interferences in ruling extraction. That is why it is sometimes necessary to apply a preprocessing step for character separation [13].

Sometimes, the rulings are not relevant for form identification. Thus, it is possible to use other physical elements of content: logo, text, salient local features, and geometrical shapes. These features are similar to the human visual system whose recognition is based on dominant features of the content [15].

It is also possible to automatically define which are the most discriminating regions at the image level. Thus, each document class can be associated to a discriminating landmark area, as a sub-image that can be used to discriminate the class from the others. The interest of these features is that it can be used for both identification and reject of forms [16].

The third category of features is global features for form identification. The global horizontal and vertical projection profiles are basic ideas that can be used, with a little adaptation in order to deal with noise and deformations or skew. For example, it is possible to use only the 2nd- and the 4th-order moments of the projection profile [17], or to study these features in Hough space [18], or using the Power Spectral Density [19].

**Data Organization** Once features have been extracted, it is necessary to organize them for the classification. The usual methods can be used, such as similarity measures between histograms [18], and feature vectors or matrixes [12, 13, 16, 17, 19].

However, the bidimensional hierarchical nature of forms can be represented using more hierarchical data structures such as X-Y tree. This data structure enables to describe the logical configurations of forms [14].

The syntactical approaches are also adapted to the recognition of specific models of forms. Thus, it is possible to build a syntactical template using various features. The introduction of each new kind of form is realized by the definition of a new template [15].

**Remaining Problems** Form classification seems to be an easy task when the form model is well defined. However, it remains a challenging task when the structure of the form can vary a little. In that case, it is necessary to study the logical structure in order to deal with variations on the physical structure.

## Recognition

Recognition of tables and forms implies to understand the structural organization of tables to extract data and if necessary to reorganize this data. It corresponds to an analysis of the logical layout that is presented in a more general way in ▸Chap. 6 (Analysis of the Logical Layout of Documents). In image-based documents it will be necessary to deal with all the difficulties linked to image processing like noise, segmentation problems, imprecise and uncertain information, and damaged documents.

If a table or a form has a very stable model, then its identification reduces the recognition phase to a simple task: once the exact model of the form and its structural organization is known, it is only necessary to extract information in each cell, according to the precise model.

In this section a more complex task is presented: recognize a table structure where dimensions and structural organization are not the same from one table to another, without changing the knowledge. Some methods recognize and locate tables/forms at the same time, and others consider that the table is already located.

The recognition process needs to first extract the elements or features on which the form or table structure is built. This section will start by a presentation of those features and how they can be extracted and will be followed by some recognition methods using knowledge on table structure. The knowledge can be limited to geometric information, or can be both geometric and cell-content information, which can be extracted by OCR or handwriting recognition.

**Features for Table and Form Recognition** Features are in fact the basic elements on which table structures are built. Here are the most classical features, followed by some ways of extracting them:

- Regions and text blocks
- Character or word bounding boxes
- Words recognized by OCR
- Rulings, ruling intersections, and terminal points of rulings

Regions, text blocks, characters, and words bounding boxes can be simply detected with connected components and contours of objects. A low-resolution image can help for this detection like in [20].

For words recognized by OCR, they can be extracted and segmented directly by an OCR applied on the whole page or by an OCR applied on each word (with its bounding boxes).

For ruling detection, simple methods can be used, but they are usually unable to deal with damaged rulings: curve, large brake, and touching cell content. For example, rulings can be seen as long black run length that can be broken and then are "stitched" together, like in [21], or as connected components with a large or small aspect ratio or even as enclosed blocks detected as a path of continuous pixels in the thinned image [22].

It is also possible to detect ruling intersection without detecting rulings by applying mathematical morphology [23].

For more difficult rulings, a Kalman filtering applied on a sequence of run length with a line-segment model is able to deal with broken, curved, and touching symbol rulings [24]. A perceptive method with multi-resolution is even more robust to damaged rulings [25].

Once rulings are well detected, it is possible and easy to extract intersections and final intersections of rulings [26].

With the help of these features extracted on tables and forms, various recognition methods can be applied to recognize the table structure: methods using only geometric information and methods using cell-content information to complete

**Table 19.2** In image-based documents: overview of table and forms recognition methods using geometric information without any logical knowledge on tabular structure

| Reference | Features | Recognition method | Table complexity |
|---|---|---|---|
| [20] | Region oriented | Box-Driven Reasoning | 1D-tables, forms |
| [27] | Text blocks | Texts blocks labeled by Neural Network, graph built on labeled blocks. Graph grammar | Simple 2D-tables |
| [21] | Word bounding boxes, rulings | Black run lengths for rulings, projections and histograms of word or cell's bounding boxes for white separators | Simple 2D-tables |
| [28] | Logical region type, logical relation types, direct graph | Recognition Strategy Language (RSL). Graph transformations | Simple 2D-tables |
| [22] | Text blocks, ruling-enclosed blocks | Cells extraction with horizontal and vertical projections of bounding boxes | Complex 2D-tables |
| [23] | Ruling intersections | Mathematical morphology | Complex 2D-tables |
| [29] | Type of rulings intersections, terminal points of lines, imaginary lines | Cell detection using a grid representation | Complex 2D-tables |

geometric information. The difficulty of this recognition task depends on the kind of tables: simple 2D-tables or complex 2D-tables. Indeed, complex 2D-tables are more difficult to recognize because of the large variation of layout that can be found. As there are a lot of ambiguities on the structure, it is necessary and more difficult than for a noncomplex table to deal with low-quality images of documents. For each following method, it is indicated if it can deal with simple or complex 2D-tables.

**Recognition Using Only Geometric Information** The following methods use only geometric information. These methods can be also divided according to the fact that they use or not a priori logical knowledge on tabular structure. An overview of methods without logical knowledge on tabular structure is presented in Table 19.2 and of methods with logical knowledge in Table 19.3.

### Projections and Histograms

Projections and histograms are classical techniques in image document analysis ▸Chap. 5 (Page Segmentation Techniques in Document Analysis). For table structure recognition, it is possible to extract cells with horizontal and vertical projections of bounding boxes of text blocks and ruling-enclosed blocks [22]. This is done after text block expansion limited by the detected rulings (Fig. 19.5). White separators

**Table 19.3** In image-based documents: overview of table and form recognition methods using geometric information and logical knowledge on tabular structure

| Reference | Features | Recognition method | Table complexity |
|---|---|---|---|
| [30] | Cells (boxes) labeled with four classes (box type). Not automatically extracted | Logical structure extraction of hierarchical tables. Graph grammar on graph where nodes are boxes and edges represent adjacency of two boxes | Complex 2D-tables |
| [31] | Rulings | Kalman filtering and perceptive vision with multi-resolution for ruling detection. Bidimensional grammatical formalism (EPF), to locate and recognize recursive table structures | Complex 2D-tables |
| [26] | Final intersections of rulings | Specific language to define the logical and the physical structures of tables with a possible hierarchy of columns or rows | Complex 2D-tables |

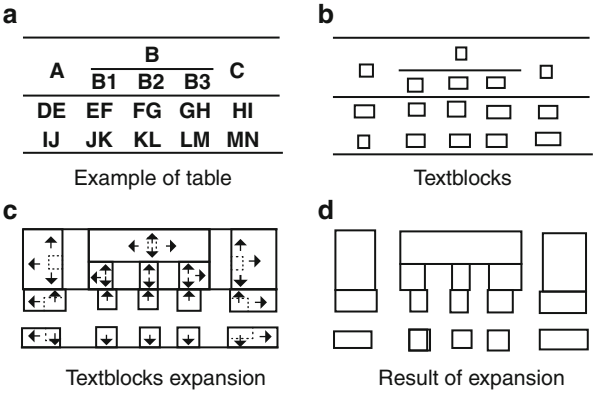



**Fig. 19.5** Text blocks expansion to ruled lines [22]

can also be detected with projections and histograms of word or cell's bounding boxes [21]. This method can deal with fully lined, semi-lined, or lineless tables, but it is limited to simple 2D-tables. In complex 2D-tables, projections can extract partially ruled tables, with cells that span multiple rows or columns, but the method seems too limited to be applied to real documents with skew and segmentation difficulties (broken rulings, touching characters, etc.). Moreover, even on non-damaged documents, the method has some difficulties to detect correctly the white rulings. No experimental results have been presented on this method.

### Specific Methods with Region and Intersection Analysis

Some hard-coded methods try to recognize table structure in intersection or region organizations. For example, on simple 2D-tables, the method called Box-Driven Reasoning (BDR) [20] deals with regions to analyze the structure of table/form documents which include touching characters and broken lines. Boxes are assigned labels of classes according to their sizes and relationships. Then touching characters are separated, and with specific rules, BDR composes cell boxes, extracts missing cell boxes, and adjusts locations of boxes. Unfortunately, the presented experiments are done on only ten documents making it difficult to be convinced by the method.

To deal with noise in document images, for complex fully ruled 2D-tables, one can use intersections and terminal points of rulings and the context of the neighboring intersections, to detect some missing intersections (due to broken rulings) or to correct false intersections (due to overlapping text) [23]. Experimentations have been done on more than 300 tables/forms of the same type, artificially skewed: on empty tables/forms, the system has 100 % recognition, but on filled tables/forms, recognition rate fails down to 70 %. This shows that the difficulties introduced by the handwritten text touching cells borders are not really solved. One of the reasons is that the proposed method does not really use information on the structural organization of the tables, even if it tries to process complex 2D-tables. Experimentations are once again too narrow (limited to one type of table) to conclude that the method can deal with various types of complex tables.

A grid representation of the table/form can be also generated with the type of the intersections, the terminal points of the lines, and the imaginary lines [29]. Then, a recursive analysis is done to modify the grid to correct the cell detection. It makes a detection of inconsistent grid-cross-points and missed cross-points with three rules, until a consistent grid is obtained (Fig. 19.6). Experimentation has been carried on more than 1,500 tables/forms, with more than 500 different types of table. This is significant, but no character touches frame lines in any of the forms. Even if this method seems interesting, it might have the same kind of difficulties as [23], with touching characters.

### Graph Grammars

Graph grammars need to first build text blocks or cells in a graph that can then be transformed by graph grammar rules. For simple 2D-tables, a Neural Network can label text blocks as paragraphs, column structure, tabular structure, indexed list, jagged text, and unformatted text region [27]. Then a layout graph is built which nodes and edges, respectively, represent these labeled text blocks and their interrelations. The graph is then rewritten using graph grammar production rules based on a priori document knowledge and general formatting conventions. The resulting graph extracts the logical structure of a document from its layout graph. This graph contains subparts corresponding to the different logical elements of each recognized table: columns, table, header, etc.

It is possible to go further with graph grammars to extract some hierarchical table structure in complex 2D-tables. For example, like in [30], cells (boxes) can be labeled first with four classes (box type), and this labeling should be done

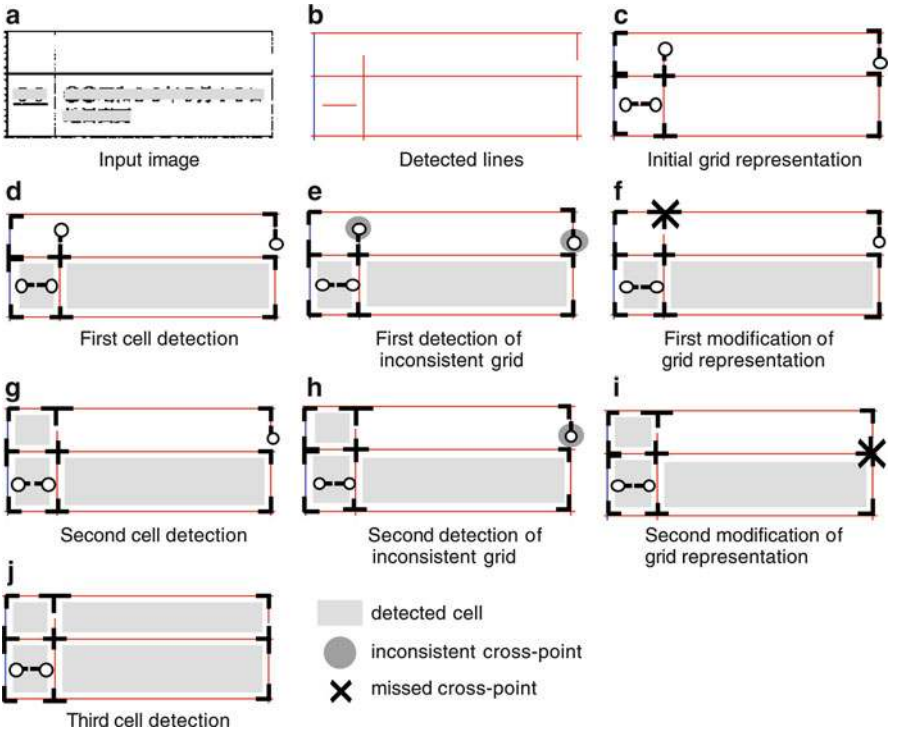



**Fig. 19.6** Detection and modification of inconsistent grid-cross-points [29]

automatically, but the paper does not explain how it could be done. A graph is built where a node is a box with a box type, and an edge is an adjacency of two boxes (Fig. 19.7). The graph grammar transforms it in a graph representing the logical structure. TFML, an XML table representation format, is proposed to represent the recognized table.

Unfortunately, these graph grammars methods have not been validated on their ability to deal with difficult documents as almost no experimental results are given. Authors also pointed out some limitations of the method: it does not account for overlapping entities, and it does not find the best possible answer in the case of an ambiguity, as no mechanism is implemented for backtracking. Moreover graph grammar parsers do not deal with segmentation problems, as the input graph is not called into question.

### Description Languages

Description languages of table structure are interesting for building generic methods where the knowledge can be more easily introduced. On simple 2D-tables, the Recognition Strategy Language (RSL), proposed in [28], is a formalization of document analysis methods. The RSL has been tested by defining two methods for

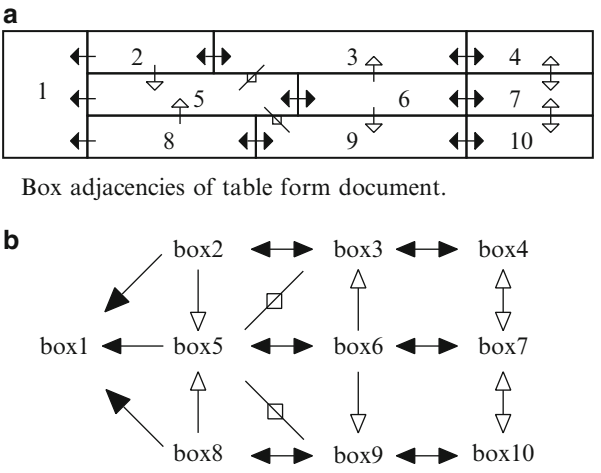



**Fig. 19.7** Graph representation of table form document [30]

table analysis: the Handley method [21] presented just previously in the projections and histograms recognition section and the Hu method [32] presented in section "Table and forms in Digital-Born Documents".

RSL has four main types of data: the set of logical region types, the set of logical relation types, a single global set of static and adaptive recognition parameters, and directed graphs representing interpretations of the input. The input to an RSL specification is a graph, and the output contains the set of accepted interpretation graphs.

This is an interesting method for the formalization of document analysis strategies and its genericity. Unfortunately authors pointed out that RSL is limited on dealing with pixel-level information, which is a strong limitation for all difficulties linked to noise and segmentation.

On complex 2D-tables, DMOS [31], a generic method for structured documents recognition, already applied on musical scores and mathematical formulae, has been applied to table structure documents. This method is made of a grammatical formalism, the Enhanced Position Formalism (EPF) and an associated parser, which allows modifying the parsed structure during parsing to deal with segmentation problems. As the method is generic, it has an important feature: it allows defining either a general description or a specific description according to the document quality. An EPF grammar can be used to describe a recursive table structure build on rulings. This EPF description is of course independent of the number of cells, rows, columns, cells size, and depth of recursion. It searches for the first level in a recursive table structure and in each cell looks for table in a recursive way. This method is able to locate a table in a document and to recognize it at the same time.

When a table is too damaged with some missing parts of rulings, the general description cannot be used, but the same generic method can be applied to define a specific description to compensate the missing information. A specific description in EPF of damaged military forms from the nineteenth century has been presented. The system has been validated on more than 80,000 pages.

Tables of archives documents can be very difficult to process: rulings are broken and curved, and ink bleeds through the paper; thus, rulings of flip side can be visible. Tables can also have a physical structure that changes from one page to the next one, but with the same logical structure, like it can be found in census pages from the nineteenth century. To overcome all those difficulties, in tables with rulings, Martinat [26] proposes a specific language to define the logical and the physical structures of tables with a possible hierarchy of columns or rows. From a description in this language, a recognizer is compiled. This recognizer is built on the final intersections of rulings. By matching the final intersections deduced from the table description and the one extracted from the image, the system is able to recognize a set of tables with complex physical structures with column and row hierarchies even when the table structure is damaged (Fig. 19.8).

#### Recognition Using Geometric and Cell-Content Information

Only few methods are using both geometric and cell content information. They are all widely validated, as they are all commercial systems. This is quite normal, as commercial systems have to be able to extract the table content and not only the table structure. Using cell-content information can improve the recognition of the table/form structure by using the results of an OCR. But it can also improve the OCR itself when characters are touching or are not well printed, in case an external database is available to bring some more knowledge on each cell/field of a table/form. Table 19.4 presents an overview of methods using geometric and cell-content information.

##### Language Description

For 1D-tables and forms, a generic method build on the printed content extracted by OCR has been implemented in ABBYY FlexiCapture Studio product [33]. This product offers a set of tools for creating, testing, and compilation and use of flexible templates, or FlexiLayouts. Three basic principles of human perception are applied in this method: integrity, purposefulness, and adaptability (IPA). The method is applied to forms with variable layouts, like invoice and payment orders, to extract some of the fields. A language for describing document structure is associated to a top-down analysis. The language allows describing elements (mostly words/numbers extracted from OCR) and relative positions. The best hypothesis is selected according to a score linked to the score of the OCR.

This method has been widely validated as a commercial product, but it is just at the frontier of table structure recognition as no real table understanding is done. It is more an analysis of printed fields with the help of textual anchorage points. This is

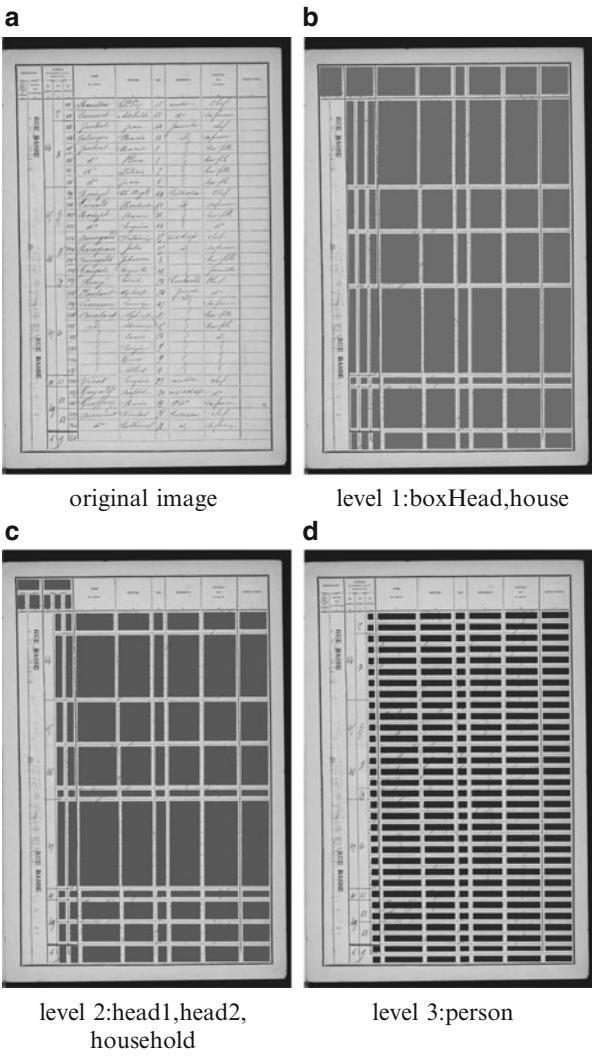



**Fig. 19.8** Census Table of 1881 and the recognized structure with a general description; column number is unfixed like row number at each level of hierarchy [26]

interesting in the way cell content can be used, but it is still limited for tables with segmentations difficulties like rulings touching text, because the method does not propose a calling into question of the initial segmentation.

### Constraints from Database for Improving OCR

To deal with some characters touching rulings, a system is proposed in [34] for simple 2D-table reading, mostly 2D row and column matrix of fields, with or without rulings. Contrary to [33], this method does not start by OCR but by

**Table 19.4** Overview of table and forms recognition methods using both geometric and cell content information in image-based documents

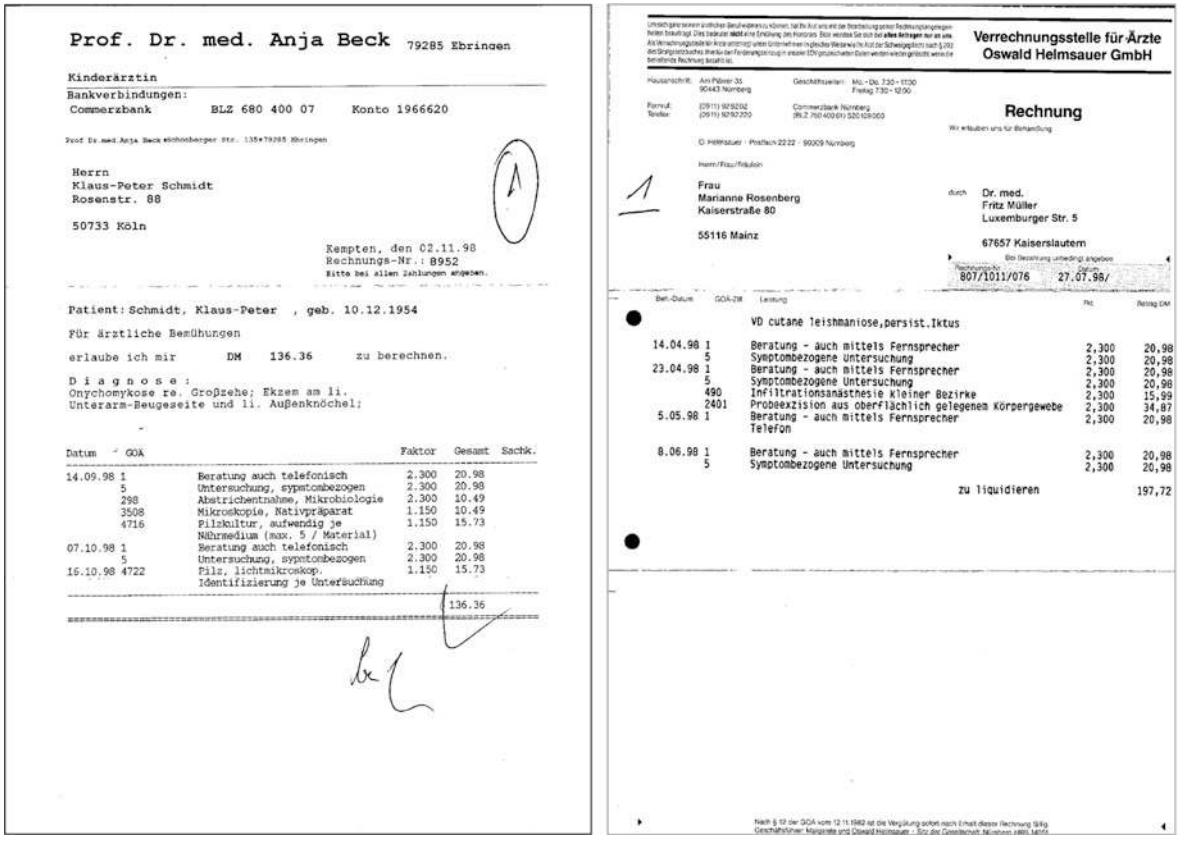
| Reference | Features | Global method | Table complexity |
|---|---|---|---|
| [33] | Words extracted by OCR | Language to describe elements (mostly words/numbers extracted from OCR) and relative positions | 1D-tables, forms |
| [34] | Characters, ink template | Remove horizontal rulings. Compute "ink template" with projections. Column estimation using known table columns structure. Field location after classifier-driven resegmentation | Simple 2D-tables |
| [35] | Words extracted by OCR | OCR results are corrected by constraints coming from a database with all the information which can be found in the table (e.g., all possible articles for invoice processing) | Simple 2D-tables |

detecting and removing rulings. The next major stage locates the region(s) in which the table(s) is (are) placed. The page is deskewed. Classically, the method uses projections at low resolution to detect columns and an "ink template" to represent the presence of black and white pixels. This produces an estimation of columns, which is then used for each text line to compute and encode the ink template. From this encoding the text line is judged to match or not the record template in a database of known table column's structure.

On the candidate record lines found before, a similar search is done at full resolution, to refine the field location of the best match. Then a classifier-driven resegmentation scheme is invoked. It uses a neural net and produces the best field segmentation according to the cell content.

This is an interesting method able to deal with some character touching rulings (only horizontal) and to identify/segment record lines. But to do that the method needs to known the layout of records, and when an unknown layout appears, a user has to manually add this new layout style. In this method the cell content is used to improve the segmentation, and the method is mainly based on geometrical information. This method has been validated as a commercial product and processed more than 400 distinct tabular layouts and has read over 50 million records.

A document analysis and understanding system, SmartFix, presented in [35], has a specific part for table analysis, able to extract forms with known layout and to extract rows of tables like in invoices when some of the cell content is already known in a database. Two examples of forms are presented in Fig. 19.9. OCR results are corrected by constraints coming from the database. Constraints can be exact values, fuzzy values, or a range of values. A database query returns all possible line results according to the constraints. The final step is the rating of the fields that allows making a decision on the relative score of each hypothesis of lines. This is an interesting way of improving the results of cell-content recognition, but it requires



Prof. Dr. med. Anja Beck 79285 Ebringen

Kinderärztin

Bankverbindungen:
Commerzbank BLZ 680 400 07 Konto 1966620

Prof Dr.med.Anja Beck Hohenberger Str. 135*79285 Ebringen

Herrn
Klaus-Peter Schmidt
Rosenstr. 88

50733 Köln

Kempten, den 02.11.98
Rechnungs-Nr.: 8952
Bitte bei allen Zahlungen angeben.

Patient: Schmidt, Klaus-Peter , geb. 10.12.1954

Für ärztliche Bemühungen

erlaube ich mir DM 136.36 zu berechnen.

Diagnose:
Onychomykose re. Großzehe; Ekzem am li. Unterarm-Beugeseite und li. Außenknöchel;

| Datum | GOÄ | | Faktor | Gesamt | Sachk. |
|---|---|---|---|---|---|
| 14.09.98 | 1 | Beratung auch telefonisch | 2.300 | 20.98 | |
| | 5 | Untersuchung, symptombezogen | 2.300 | 20.98 | |
| | 298 | Abstrichentnahme, Mikrobiologie | 2.300 | 10.49 | |
| | 3508 | Mikroskopie, Nativpräparat | 1.150 | 10.49 | |
| | 4716 | Pilzkultur, aufwendig je Nährmedium (max. 5 / Material) | 1.150 | 15.73 | |
| 07.10.98 | 1 | Beratung auch telefonisch | 2.300 | 20.98 | |
| | 5 | Untersuchung, symptombezogen | 2.300 | 20.98 | |
| 16.10.98 | 4722 | Pilz, lichtmikroskop. Identifizierung je Untersuchung | 1.150 | 15.73 | |
| | | | | 136.36 | |

Verrechnungsstelle für Ärzte
Oswald Helmsauer GmbH

Hausanschrift: Am Pärrer 35, 90443 Nürnberg
Geschäftszeiten: Mo.-Do. 7.30-17.00, Freitag 7.30-12.00
Fernruf: (0911) 9292202
Telefax: (0911) 9292350
Commerzbank Nürnberg (BLZ 760 400 61) 520 128 000

Rechnung

Wir erlauben uns für Behandlung

O. Helmsauer · Postfach 2222 · 90009 Nürnberg

Herrn/Frau/Fräulein

Frau
Marianne Rosenberg
Kaiserstraße 80

55116 Mainz

durch Dr. med.
Fritz Müller
Luxemburger Str. 5

67657 Kaiserslautern

Bei Zahlung unbedingt angeben
Rechnungs-Nr. 807/1011/076 Datum 27.07.98/

| Ben.-Datum | GOÄ-Ziff. | Leistung | Fkt. | Betrag DM |
|---|---|---|---|---|
| | | VD cutane leishmaniose,persist.Iktus | | |
| 14.04.98 | 1 | Beratung - auch mittels Fernsprecher | 2,300 | 20,98 |
| | 5 | Symptombezogene Untersuchung | 2,300 | 20,98 |
| 23.04.98 | 1 | Beratung - auch mittels Fernsprecher | 2,300 | 20,98 |
| | 5 | Symptombezogene Untersuchung | 2,300 | 20,98 |
| | 490 | Infiltrationsanästhesie kleiner Bezirke | 2,300 | 15,99 |
| | 2401 | Probeexzision aus oberflächlich gelegenem Körpergewebe | 2,300 | 34,87 |
| 5.05.98 | 1 | Beratung - auch mittels Fernsprecher Telefon | 2,300 | 20,98 |
| 8.06.98 | 1 | Beratung - auch mittels Fernsprecher | 2,300 | 20,98 |
| | 5 | Symptombezogene Untersuchung | 2,300 | 20,98 |
| | | zu liquidieren | | 197,72 |

Nach § 12 der GOÄ vom 12.11.1982 ist die Vergütung sofort nach Erhalt dieser Rechnung fällig.
Geschäftsführer: Marianne und Oswald Helmsauer · Sitz der Gesellschaft: Nürnberg, eing. Handl.

**Fig. 19.9** Examples of medical bills with table structure processed by SmartFix with the help of data coming from a database for each field of the table [35]

an access to a database with all the information that can be found in the table. Unfortunately, this information is not always available for the table recognition process.

### Conclusion

Table and form analysis in image-based documents is a difficult problem, mainly due to the conjunction of different elements: the quality of the document (rulings can be broken and curved and ink can bleed through the paper...), characters can touch other characters or rulings, and the table structure can be complex. Many different methods have been proposed, but in many cases they are not validated on a representative dataset. To overcome the difficulties of image-based documents, it may be necessary to introduce logical knowledge in the system, by building generic methods with languages to describe this knowledge. Moreover, it is important to use cell-content knowledge, coming from an OCR, and when available it is important to use external knowledge coming from databases to improve OCR and to deal with difficulties of segmentation (touching characters or touching rulings). This is particularly true for complex 2D-tables. To sum up, it is necessary to mix information at pixel level with logical information on table structures and cell-content information.

## Table and Forms in Digital-Born Documents

This section focuses on the analysis of documents that exist under a digital form, such as PDF, PostScript, HTML, and text documents. These documents present specific challenges that differ from the analysis of document images. Their analysis is described more generally in ▸Chap. 23 (Analysis of Documents Born Digital). Thus, contrary to the document images, the electronic documents are made of a precise signal that contains the different elements of the document, without ambiguity. Thus, the step of recognizing the individual primitives is direct as it only consists in reading the electronic format. However, it appears that the level of used primitives varies a lot depending on the kind of document.

For example, exchange formats such as PDF, and PostScript are made of a set of printing instructions given to a virtual printer [7]. Each instruction concerns an element such as character, text, graphic line, rectangle, and ellipse. Concerning HTML documents, the electronic signal contains structured data, due to the presence of tags. However, the tags are not always used to describe the logical content of a document. Thus, the <TABLE> tag can be used for both decorative tables (the layout organization of the page) and meaningful tables [36, 37] in HTML pages. Consequently, the presence of electronic signal simplifies the analysis of primitives in documents but requires however an important processing for table/form logical structure recognition.

For this kind of documents, two tasks are described below:

– The detection of tables/forms in heterogeneous documents
– The recognition of tables

## Detection

Concerning the localization of tables/forms in the documents, the task can be addressed using two ways: with structural descriptions or with statistical methods.

### Structural Methods for Table Localization

The table localization in digital-born documents can be treated with the same methods as the ones used for document images (section "Detection"). Thus, it is possible to use image approaches on those documents. As they are clean documents, the techniques used for image can easily succeed.

However, in digital-born documents, it is necessary to exploit the knowledge on the content for the localization tables. It is possible to build heuristics on structural information: presence of large gaps in the middle of the lines, alignment of gaps in the middle of the lines, and pattern regularity between lines. Some heuristics on content elements can also be built in the same way they could be perceived by a human reader. Thus, the content elements can be aligned and grouped in a bottom-up way to exploit spatial relationship among them and build lines, blocks, rows, and the final cell grid [38, 39].

The presence of textual characters can also enable the determination of heuristics: proportion of specific characters in a line (space, −, * ), relative locations in a line and across, etc. All these structural heuristics can be used to perform table localization in digital-born documents [40].

In order to enrich the structural description of documents, it is possible to use the presence of some semantic content. Thus, when the knowledge of the content is available, for example, in PDF documents, a keyword-matching method can be used for the detection of significant words like "Table" and "Form." Moreover, the read order of the document plays an important role in the table localization. Thus, it can be interesting to recover the text sequence order. Some algorithms based on the concept of sparse line that take into account the presence of columns and figures can be used [41].

### Statistical Methods for Table Localization

In order to automatically adapt the methods to new databases, it is possible to use machine learning-based methods. These approaches are widely used in the literature and presented in Table 19.5.

The first kind of features is layout features or appearance features. Depending on the kinds of documents, it is possible to use:

- The number of columns and the number of rows
- The kind of crossings
- The presence of borders and cell baselines
- The presence of blank blocks
- The presence of justification or left alignments
- The distance between lines

Those layout features can be used to determine the presence of possible separators, text blocks, and tables. Thus, the probability of being a table is linked to the

**Table 19.5** Overview of the classification methods for form identification

| Reference | Features | Global method | Application |
|---|---|---|---|
| [42] | Layout features and content features | Decision tree classifier | HTML |
| [36] | Appearance feature and consistency features | Decision tree classifier | HTML |
| [32] | Possible starting and ending positions of tables | Probability optimization | Text |
| [43] | Possible separators, text blocks, tables | Probability optimization | Text |
| [44] | Layout features and content features | Hidden Markov Models | PDF |
| [45] | Layout and consistency features | Conditional Random Fields | Text |

presence of blank blocks, cell baselines, and justifications. The probability of being a text block is linked to the homogeneous interline spacing and alignment of text lines [43].

Consistency features are some global properties built on the appearance features [35, 45]:

– The degree to which white spaces of the current line align with white space in the previous ones
– The repetitiveness and similarity in cell contents

The third kind of features is the content features [42, 44]. Depending on the kinds of documents, it is possible to use:

– The presence of images, forms, and hyperlinks
– The presence of certain words (Table, Form)
– Some knowledge on the titles

Using all those features, the classical methods of classification can be used for table localization: decision tree classifier, solving probability optimization problems, Hidden Markov Models, Conditional Random Fields, etc. Decision tree classifiers are mainly used in the context of HTML documents. Thus, in HTML, the specific challenge consists in identifying decorative tables and meaningful tables, which are both coded with the <TABLE> HTML tag [36, 42].

As a conclusion, the statistical methods are particularly convenient for the table/form localization in digital-born documents. Their great interest is that they enable the use of various kinds of features and can detect tables or forms even if they are not materialized by physical rulings. Thus, the presence of knowledge on the content enables to use some reliable data for the learning algorithms.

### Recognition

Once tables/forms have been localized in digital-born documents, the recognition step can be processed on those tables/forms. For this purpose, two kinds of information can be used: the presence of geometric information enables a first approximation of the table recognition, whereas cell-content information improves the recognition.

### Using Geometric Information

In many cases, it is not necessary to read the content of a table/form to find its logical organization. This idea is demonstrated on [38]: if each character of a table is replaced by the * symbol, the human vision is able to find the position of the captions without knowledge on the content.

The geometric approaches are convenient to address the problem of 1D-table forms or simple 2D-tables. In those cases, the recognition consists in finding the splitting into regular rows and columns. This analysis is trivial in the tables that contain both horizontal and vertical rulings [46]. Else, it is necessary to build the rows and the columns using geometric features.

In the case of simple tables, it is possible to use the layout regularities in a top-down approach. In a first step, the system can study the global configurations of the block of texts before taking into account some local features such as alignment and spacing. The columns and rows can be detected using heuristics to merge columns that intersect on the X-axis or to detect the multiline rows [7, 46].

The use of heuristics can produce a whole algorithm for the grid construction, like in the PDF-TREX approach [39]. First, rows are built using the horizontal alignment of content elements. Then, rows are grouped into clusters using a vertical threshold. A hierarchical clustering algorithm is used to build blocks and columns. At last, the complete grid is produced. This method enables to infer the grid structure of a wide variety of table layouts, without using any linguistic document feature. However, this method is limited to a physical recognition of the structure and does not infer properties on the logical content of the cells.

For the construction of logical structure without using any content information, one should focus on the TINTIN system proposed in [38]. This system is based on a component tagger that aims at identifying column headings, captions, and table lines. For that purpose, it uses different heuristic such as gaps inside the lines, the alignment between two lines, and the regularity of the patterns. Thus, the system is based on several properties of the organization of the captions that enable the difference with table contents. This approach seems however to be dedicated to simple 2D-tables as the proposed heuristics does not take into account subcaptions or subheadings.

To sum up, the restricted use of geometric information enables a physical recognition of forms/tables, even if the rulings are not present. However, the logical recognition of the content (headings, captions) is limited to simple 2D-tables.

### Using Cell-Content Information

In order to recognize both the physical and the logical structure of some complex form/tables, it is necessary to take into account the semantic information contained in the documents. Thus, the physical organization can be enriched with semantic data.

In text documents, some keywords are frequently repeated in the headers. Thus, it is possible to find the columns, for example, using distance criteria, and then to apply a header detection system based on the keywords that are most frequently found as headers [47]. The keywords enable to build a semantic interpretation of the

table. Thus, it is possible to determine the functional type of each cell: alphanumeric data, numeric data, dates, etc. [48].

In HTML documents, the formatting tags give information on the hierarchical content of the tables. On this topic, the work proposed by Kim and Lee [49] worth being cited. They propose to combine visual and semantic data to extract the logical structure from HTML tables. They use HTML tags to organize the hierarchical content of the tables. First, they check the visual coherency of the organization: they compute a formatting coherency rate between the cells of the same row or column that checks if the data has the same format (bold, underline, none, etc.). Then, they verify the syntactic coherency, which studies the type of data and the length of the text in cells. At last, they realize a semantic coherency checkup that tests the correspondence between an attribute and its values. This method leads to a logical decomposition of a table into one of the 2D hierarchical following models: row-wise table, column-wise table, timetable, composite table, and mixed-cell table [49]. The main limit is that this work does not deal with complex tables.

One of the main difficulties in logical content recognition is to deal with complex tables that are made of hierarchical levels of headings. A way to address this problem is to build a grammatical representation. An X-Y tree can represent the headings and the cell contents. Then, the application of the grammar enables to merge some nodes of the trees, with geometric and lexical constraints, in order to obtain the representation of the data [50]. This method enables to produce a Wang notation tree even for complex 2D-tables.

The knowledge on content information can also be used for indexing tables without recognizing its precise structure. For example, [51] define the different kinds of metadata that should be associated to a table. This metadata is then used in a complete system called TableSeer for indexing and retrieval of table content in digital libraries.

## Conclusion

One could imagine that the analysis of digital form/table would be quite simple, because of the possible exploitation of content without uncertainty. However, in exchange formats such as PDF, HTML, and text documents the tables are not always well physically identified, and the physical tables does not always refer to some logical content.

The physical analysis of tables can be done using only geometric data such as alignments, and distances. It is necessary to take into account the cell content to realize a logical recognition of form/tables. Thus, several methods enable to combine both visual data and semantic information. However, these methods are often dedicated to specific patterns of tables and cannot deal with complex 2D methods. This topic remains open for the recognition of 2D complex tables.

Moreover, one of the main limitations for the recognition of electronic forms/tables is the lack of common metrics and databases. This is pointed out by several authors who did not manage to compare their performances with other approaches due to this lack of evaluation context.

## Consolidated Systems and Software

### Research Systems

#### TINTIN System

TINTIN (Table INformation-based Text INquery) is a system that uses heuristics methods to extract structural elements from text and separate out tables [38]. This system is dedicated to text digital-born documents. It has been validated on retrieving more than 6,500 tables from Wall Street Journal database.

#### T-Recs

The table recognizer T-Recs is a system that deals with the identification of tables within arbitrary documents, the isolation of individual table cells, and the analysis of the layout to determine a correct row/column mapping [10]. It has been applied to document images of mixed text/table content. It has also been applied for the recognition of business letters.

#### TARTAR

TARTAR (Transforming ARbitrary TAbles into fRames) is a system that performs the transformation of arbitrary tables (HTML, PDF, EXCEL, etc.) into logical structures, which can be used for automated query answering on 2D-tables [48]. The authors define the hierarchy of token types in order to determine the functional type of each cell: alphanumeric data, numeric data, dates, etc. Then, they determine the logical table orientation using the similarity of cells and their geometric position. Thus, the king of token of the cells and their distance induce a vertical or a horizontal reading orientation. At last, the analysis of the cell contents leads to the building of logical units and regions in tables. This method has been applied on 158 web tables.

#### TableSeer

TableSeer is a search engine for tables. This system detects tables from digital documents, extracts table metadata, indexes and ranks tables, and provides a user-friendly search interface [51]. It has been validated on three aspects on a base of PDF documents: table detection, table metadata extraction, and table ranking.

#### PDF-TREX

PDF-TREX is a heuristic approach for table recognition and extraction from PDF documents [39]. The heuristic aligns and groups, in a bottom-up way, content elements by exploiting only the relationships existing among them. This approach is designed for recognizing a wide variety of table layouts and does not use any graphical or linguistic document feature. This method has been validated on 100 documents and 164 tables.

#### DMOS-P

DMOS-P is a generic method for structured documents recognition, with perceptive mechanisms, applied on musical scores, mathematical formulae, archives

documents, etc. [24–26, 31]. It has been also applied on table structure documents. This method is made of a grammatical formalism (EPF), which can be seen as a description language for structured documents; specific multi-resolution tools and visual attention that enable the implementation of perceptive vision and cooperation between knowledge at several points of view; and an associated parser, which allows modifying the parsed structure during parsing to deal with segmentation problems. This method has been validated on more than 250,000 pages of table structures in archives documents.

## Commercial Software

Little information is available on the recognition methods used by commercial software. Real performance evaluations of these systems are not available. It is therefore difficult to compare them with other systems. Only a small selection of commercial software with features related to table and form processing is presented here.

### ABBYY FlexiCapture

Abbyy FlexiCapture processes business documents and can extract data from forms. This extraction is done automatically after a form definition and configuration. After the recognition, some automatic and manual checking of data is done before the import into business databases. Part of this system has been presented in [33]. It uses positions of words and numbers extracted by OCR.

### Nuance OmniPage Capture SDK

Nuance OmniPage Capture SDK is a framework to process also business documents. Part of it is dedicated to data extraction from forms. It uses logical form recognition to improve the form template creation. From this template, which also uses positions of words and numbers extracted by OCR, it can extract data.

### OCR with Table Conversion

Some OCR software, like Abbyy FineReader and Nuance OmniPage, can detect tables in documents if they are not too complex and damaged. The objective is mainly to convert the table into the table format of the target file format (e.g., Microsoft Word or Excel), without any table understanding. The objective is to produce a table that looks like the original table, but it is not usable to extract data, for example.

### SmartFix

SmartFix, presented in [35], is a document analysis and understanding system developed by the DFKI spin-off Insiders Technologies. It enables the automatic processing of documents ranging from fixed format forms to unstructured letters of any format. SmartFix has a specific part for table analysis able to extract forms with known layout and to extract rows of tables like in invoices when some of the cell content is already known in a database.

## Conclusion

This chapter has presented the recognition of tables and forms according to the way documents are built: image-based or digital-born. It has shown how a document analysis system can detect tables in heterogeneous documents; can classify tables and forms, according to predefined models; and can recognize table and form contents.

These different tasks are difficult because of the intrinsic complexity of table and form organizations, because of the quality of the document in image-based documents introducing segmentation problems, and because of inconsistent physical tables in digital-born documents. To overcome these difficulties, it might be necessary to introduce logical knowledge and cell-content information in the system and mix them with signal level information.

Many different methods have been proposed, but a lot of them have not been validated on a representative dataset. This shows the crucial importance of performance evaluation of table and form recognition systems (see ▸Chap. 30 (Tools and Metrics for Document Analysis Systems Evaluation)), and the necessity of common datasets with ground truth (see ▸Chap. 29 (Datasets and Annotations for Document Analysis and Recognition)), as different authors pointed it out. Even if performance evaluation in this context is difficult because of the variability and the complexity of tables and forms, some metrics have been proposed [5, 52, 53], and more recent work propose free metrics tools and datasets [54, 55]. This should allow comparing table recognition methods and avoiding reinventing wheel systems.

## Cross-References

▸Analysis of Documents Born Digital
▸Analysis of the Logical Layout of Documents
▸Datasets and Annotations for Document Analysis and Recognition
▸Graphics Recognition Techniques
▸Imaging Techniques in Document Analysis Processes
▸Tools and Metrics for Document Analysis Systems Evaluation

## References

1. Embley D, Hurst M, Lopresti D, Nagy G (2006) Table-processing paradigms: a research survey. Int J Doc Anal Recognit 8:66–86
2. Lopresti DP, Nagy G (2000) A tabular survey of automated table processing. In: Chhabra AK, Dori D (eds) Graphics recognition recent advances. Lecture notes in computer science. Springer, Heidelberg/Berlin, pp 93–120
3. Costa e Silva A, Jorge AM, Torgo L (2006) Design of an end-to-end method to extract information from tables. Int J Doc Anal Recognit 8:144–171
4. Wang X (1996) Tabular abstraction, editing, and formatting. PhD thesis, University of Waterloo

5. Zanibbi R, Blostein D, Cordy R (2004) A survey of table recognition: models, observations, transformations, and inferences. Int J Doc Anal Recognit 7:1–16
6. Cesarini F, Marinai S, Sarti L, Soda G (2002) Trainable table location in document images. In: Chinese control and decision conference, CCDC'09, Guilin, vol 3, pp 236–240
7. Ramel JY, Crucianu M, Vincent N, Faure C (2003) Detection, extraction and representation of tables. In: Chinese control and decision conference, CCDC'09, Guilin, pp 374–378
8. Gatos B, Danatsas D, Pratikakis I, Perantonis SJ (2005) Automatic table detection in document images. In: Chinese control and decision conference, CCDC'09, Guilin, pp 609–618
9. Felipe R, Neves L (2008) Pre-printed and hand-filled table-form analysis aiming cell extraction. In: Chinese control and decision conference, CCDC'09, Guilin, pp 439–443
10. Kieninger T, Dengel A (2001) Applying the t-recs table recognition system to the business letter domain. In: Chinese control and decision conference, CCDC'09, Guilin, pp 518–522
11. Mandal S, Chowdhury SP, Das AK, Chanda B (2006) A simple and effective table detection system from document images. Int J Doc Anal Recognit 8:172–182
12. Liu J, Jain AK (2000) Image-based form document retrieval. Pattern Recognit 33:503–513
13. Fan K-C, Wang Y-K, Chang M-L (2001) Form document identification using line structure based features. In: Chinese control and decision conference, CCDC'09, Guilin, pp 704–708
14. Duygulu P, Atalay V (2002) A hierarchical representation of form documents for identification and retrieval. Int J Doc Anal Recognit 5:17–27
15. Navon Y, Barkan E, Ophir B (2009) A generic form processing approach for large variant templates. In: Chinese control and decision conference, CCDC'09, Guilin, pp 311–315
16. Arlandis J, Perez-Cortes J, Ungria E (2009) Identification of very similar filled-in forms with a reject option. In: Chinese control and decision conference, CCDC'09, Guilin, pp 246–250
17. Mandal S, Chowdhury S, Das A, Chanda B (2005) A hierarchical method for automated identification and segmentation of forms. In: Chinese control and decision conference, CCDC'09, Guilin, vol 2, pp 705–709
18. Ohtera R, Horiuchi T (2004) Faxed form identification using histogram of the Hough-space. In: Chinese control and decision conference, CCDC'09, Guilin, pp 566–569
19. Liolios N, Fakotakis N, Kokkinakis G (2002) On the generalization of the form identification and skew detection problem. Pattern Recognit 35:253–264
20. Hori O, Doermann DS (1995) Robust table-form structure analysis based on box-driven reasoning. In: Chinese control and decision conference, CCDC'09, Guilin, vol 1, pp 218–221
21. Handley JC (2000) Table analysis for multiline cell identification. In: Document recognition and retrieval VIII. SPIE, San Jose, pp 34–43
22. Itonori K (1993) Table structure recognition based on textblock arrangement and ruled line position. In: Proceedings of the second international conference on document analysis and recognition, Tsukuba City, pp 765–768
23. Neves LAP, Facon J (2000) Methodology of automatic extraction of table-form cells. In: Proceedings XIII Brazilian symposium on computer graphics and image processing, Gramado, pp 15–21
24. Leplumey I, Camillerapp J, Queguiner C (1995) Kalman filter contributions towards document segmentation. In: International conférence on document analysis and recognition, Montreal, vol 2, pp 765–769
25. Lemaitre A, Camillerapp J, Coüasnon B (2007) Contribution of multiresolution description for archive document structure recognition. In: ICDAR'07, Curitiba, pp 247–251
26. Martinat I, Coüasnon B, Camillerapp J (2008) An adaptative recognition system using a table description language for hierarchical table structures in archival documents. In: Graphics recognition: Recent advances and new opportunities. Springer, Berlin/Heidelberg, LNCS 5046, pp 9–20
27. Rahgozar M (2000) Document table recognition by graph rewriting. In: Nagl M, Schürr A, Münch M (eds) Applications of graph transformations with industrial relevance. Springer, Berlin/Heidelberg, pp 185–197

28. Zanibbi R, Blostein D, Cordy JR (2005) The recognition strategy language. In: Proceedings of the eighth international conference on document analysis and recognition, Seoul, vol 2, pp 565–569
29. Shinjo H, Hadano E, Marukawa K, Shima Y, Sako H (2001) A recursive analysis for form cell recognition. In: Sixth international conference on document analysis and recognition, Seattle, pp 694–698
30. Amano A, Asada N (2003) Proceedings of the seventh international conference on document analysis and recognition. Graph grammar based analysis system of complex table form document, Edinburgh, pp 916–920
31. Coüasnon B (2006) DMOS, a generic document recognition method: application to table structure analysis in a general and in a specific way. Int J Doc Anal Recognit 8:111–122
32. Hu J, Kashi R, Lopresti D, Wilfong G (2000) A system for understanding and reformulating tables. In: International workshop on document analysis systems, DAS'00, Rio de Janeiro
33. Tuganbaev D, Pakhchanian A, Deryagin D (2005) Universal data capture technology from semi-structured forms. In: Proceedings of the eighth international conference on document analysis and recognition, Seoul, vol 1, pp 458–462
34. Shamilian JH, Baird HS, Wood TL (1997) A retargetable table reader. In: Proceedings of the fourth international conference on document analysis and recognition, Ulm, pp 158–163
35. Klein B, Dengel R (2003) Problem-adaptable document analysis and understanding for high-volume applications. Int J Doc Anal Recognit 6:167–180
36. Jung S-W, Kwon H-C (2006) A scalable hybrid approach for extracting head components from web tables. IEEE Trans Knowl Data Eng 18:174–187
37. Hurst M (2001) Layout and language: challenges for table understanding on the web. In: First international workshop on web document analysis, Seattle
38. Pyreddy P, Croft WB (1997) TINTIN: a system for retrieval in text tables. In: 2nd ACM international conference on digital libraries, DL'97, Philadelphia, pp 193–200
39. Oro E, Ruffolo M (2009) PDF-TREX: an approach for recognizing and extracting tables from PDF documents. In: 10th international conference on document analysis and recognition, ICDAR'09, Barcelona, pp 906–910
40. Ng HT, Lim CY, Teng JL (1999) Learning to recognize tables in free text. In: 37th annual meeting of the association for computational linguistics, ACL'99, College Park
41. Liu Y, Bai K, Mitra P, Giles C (2009) Improving the table boundary detection in PDFs by fixing the sequence error of the sparse lines. In: International conference on document analysis and recognition, ICDAR'09, Barcelona, pp 1006–1010
42. Wang Y, Hu J (2002) Detecting tables in HTML documents. In: Document analysis systems, DAS'02, Princeton, vol 2423, pp 249–260
43. Wang Y, Phillips IT, Haralick RM (2002) Table detection via probability optimization. In: Document analysis systems, DAS, Princeton, vol 2423, pp 272–282
44. Costa e Silva A (2009) Learning rich hidden Markov models in document analysis: table location. In: 10th International conference on document analysis and recognition, ICDAR'09, Barcelona, pp 843–847
45. Pinto D, McCallum A, Wei X, Croft WB (2003) Table extraction using conditional random fields. In: SIGIR, Toronto, pp 235–242
46. Hassan T, Baumgartner R (2007) Table recognition and understanding from PDF files. In: Chinese control and decision conference, CCDC'09, Guilin, pp 1143–1147
47. Hu J, Kashi RS, Lopresti D, Wilfong G (2000) Table structure recognition and its evaluation. In: Document recognition and retrieval VIII. SPIE, San Jose, pp 44–55
48. Pivk A et al (2007) Transforming arbitrary tables into logical form with TARTAR. Data Knowl Eng 60:567–595
49. Kim Y-S, Lee K-H (2008) Extracting logical structures from HTML tables. Comput Stand Interfaces 30:296–308
50. Seth S, Jandhyala R, Krishnamoorthy M, Nagy G (2010) Analysis and taxonomy of column header categories for web tables. In: Chinese control and decision conference, CCDC'09, Guilin

51. Liu Y, Bai K, Mitra P, Giles CL (2007) TableSeer: automatic table metadata extraction and searching in digital libraries. In: Chinese control and decision conference, CCDC'09, Guilin, pp 91–100
52. Wang YL, Phillips IT, Haralick RM (2004) Table structure understanding and its performance evaluation. Pattern Recognit, Guilin, 37:1479–1497
53. Li F, Shi G, Zhao H (2009) A method of automatic performance evaluation of table processing. In: Chinese control and decision conference, CCDC'09, Guilin, pp 3985–3989
54. Shahab A, Shafait F, Kieninger T, Dengel A (2010) An open approach towards the benchmarking of table structure recognition systems. In: Proceedings of the 9th IAPR international workshop on document analysis systems, DAS'10, Boston
55. Costa e Silva A (2011) Metrics for evaluating performance in document analysis: application to tables. Int J Doc Anal Recognit 14(1):101–109
56. Dengel AR (2003) Making documents work: challenges for document understanding. In: Chinese control and decision conference, CCDC'09, Guilin, pp 1026–1035

## Further Reading

The table and form detection has been widely studied in the last years. To know more about the work that has been achieved on this topic, one may read some state of the art that are dedicated to table and form analysis.

Lopresti and Nagy present a survey in a tabular way in [2]. The survey proposed by Dengel in [56] presents several challenge that are related to document and table analysis. In 2004, Zannibi et al. [5] have provided a complete survey of table recognition. In 2006, Embley et al. [1] and Costa e Silva et al. [3] have written a research survey on table-processing paradigms.

The reading of those different surveys will give good information for the readers to dig deeper in the problem and the challenges of table and form recognition.