Datasets and Annotations for Document Analysis and Recognition

29

Ernest Valveny

Contents

Abstract

The definition of standard frameworks for performance evaluation is a key issue in order to advance the state-of-the-art in any field of document analysis since it permits a fair and objective comparison of different proposed methods under a common scenario. For that reason, a large number of public datasets have emerged in the last years. However, several challenges must be considered when creating such datasets in order to get a sufficiently large collection of

E. Valveny
Departament de Ciències de la Computació, Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain
e-mail: ernest@cvc.uab.es; Ernest.Valveny@uab.cat

D. Doermann, K. Tombre (eds.), *Handbook of Document Image Processing and Recognition*, DOI 10.1007/978-0-85729-859-1_32,

representative data that can be easily exploited by the researchers. In this chapter we review different approaches followed by the document analysis community to address some of these challenges, such as the collection of representative data, its annotation with ground-truth information, or the representation using accepted and common formats. We also provide a comprehensive list of existing public datasets for each of the different areas of document analysis.

Keywords
Benchmarking • Generation of synthetic data • Ground-truthing • Performance evaluation • Public datasets

Introduction

Nowadays, performance evaluation using standard datasets is a common practice in document analysis. This is the only way to do a fair comparison of existing algorithms as has been remarked in several chapters of this book. However, the creation of such standard datasets can be a costly process in terms of data selection, acquisition, and annotation. Thus, throughout the years the research community has devoted a big effort not only to the generation of public collections of data but also to the development of tools and frameworks that could make this process much easier.

The first works on standard datasets related to document analysis date back to the early 1990s mainly in the area of OCR and zone segmentation motivated by the need of "common data sets on which to develop and compare the performance of the algorithms" [50] and the evidence that "it is time for a series of comprehensive standard document databases to be constructed and made them available to researchers" [50]. This line of research was rapidly extended to other applications of document analysis, such as binarization, vectorization, symbol recognition, table recognition, and signature verification, leading to a large number of public datasets that will be presented and discussed later in this chapter. Probably, the best evidence of this effort is the increasing number of competitions related to different areas of document analysis that are being organized in the framework of a number of international conferences and workshops.

Although every particular field of document analysis has its own properties, there are some common issues to be taken into account in the design and creation of any dataset independently of the domain of application. They are mainly related to the selection, acquisition, and annotation of data and to the different ways of storing information and organizing the dataset. In the next section we will briefly introduce them. They will be discussed more deeply in the rest of the chapter. In particular, section "Data Collection and Annotation" is devoted to the selection, collection, and annotation of data. This is probably the most costly process (in terms of time and human resources) in the creation of any dataset. Therefore, a number of protocols, frameworks, and tools have been proposed to alleviate this effort. They include efficient ways to label either manually or semiautomatically large amounts of real data. An alternative is the generation of synthetic data. In this case, it is

necessary that the definition of models of degradation is able to generate artificial data as similar as possible to real data. Later, section "Formats and Standards" deals with the representation and storage of data (both input data and annotation data), making emphasis on those efforts undertaken to establish standard formats. Section "Public Databases for Document Analysis and Recognition" will include a comprehensive list of existing public datasets grouped by domain of application. It will contain a brief description of the main characteristics of each dataset and links on how to obtain them. Additionally, some specific issues that can concern the creation of datasets for each particular domain, such as ground-truth information, the organization of the datasets, or the protocol of evaluation, will also be discussed. Finally, section "Summary" will summarize the main ideas presented throughout the chapter, drawing the main conclusions about the current state of development of standard datasets for document analysis and also pointing out some challenges for future work. Finally, it is worth noticing that the creation of standard datasets is only one of the necessary elements for benchmarking and evaluating document analysis algorithms. The other required elements, mainly evaluation metrics and protocols, will be discussed in ▸Chap. 30 (Tools and Metrics for Document Analysis Systems Evaluation).

General Issues in the Design and Creation of Datasets

A set of general issues should be considered in the process of generation of any standard dataset for performance evaluation. They cover all steps necessary for the creation of a dataset: data collection, ground-truth definition, and data representation, storage, and organization. In the following they will be briefly introduced, being further discussed in the next sections of the chapter.

Selection of Data: A requirement of any dataset intended to be used in performance evaluation is being realistic and representative of the set of images encountered in real-world applications. Being realistic implies that the set of images included in the dataset should be as similar as possible to real images. Obviously this is easy to achieve if the dataset contains images directly taken from real applications. However, as it will be further discussed later in this section, the cost of acquisition and annotation of real data sometimes forces the use of synthetic data. In these cases, realism of the dataset relies on the existence of accurate models of data generation – see section "Generation of Synthetic Data."

Being representative means that the dataset should contain a balanced mix of all the classes of documents or entities that exist in a given application domain. How to achieve this balance depends very much on the application. For instance, in datasets for page segmentation, the emphasis should be put in collecting pages from many different types of documents including different combinations, layouts and formats of text, graphics, figures, tables, and photographs. However, in a different domain such as word recognition, the importance will be on issues such as having a realistic balance of all the words in the vocabulary or including

multiple handwriting styles through the enrollment of different writers. In this way, for every domain a set of important properties can be defined. This will be illustrated in section "Public Databases for Document Analysis and Recognition" where a description of the data included in existing datasets will be provided. The existence of multiple datasets for a given domain can be seen as the consequence of the difficulty of gathering all the types of data in a single dataset. Thus, in some cases, multiple datasets can be seen as complementary (more than redundant) in order to get a complete set of representative images.

The presence of noise and distortion is intrinsic to document processing. Thus, the creation of datasets should also guarantee the inclusion of realistic and representative degraded images, that is, images with a right balance of types and levels of degradation similar to those encountered in the real world. Sources of degradation and distortion can be very diverse, depending on the application domain: acquisition noise, manipulation, aging effects, handwriting variability, different layouts of the objects in the document, etc. However, sometimes it can also be interesting to generate challenging datasets with extreme levels of noise, higher than in real images, in order to push the limits of existing methods further away.

Data Acquisition/Generation: Once decided which kind of data will be included in a dataset, the next step is collecting a sufficiently large set of data according to those requirements. What a sufficiently large means is not easy to determine. It will also depend on each application domain and will rely on the already discussed properties of being realistic and representative. Collecting and annotating large sets of data is an extremely costly task. Thus, in some cases, dataset developers have decided to generate data in a synthetic way using some model of distortion applied to clean images. Although this makes collection and annotation of data much easier, it can go against the property of realism of data. Advantages and drawbacks of real vs. synthetic data, as well as several models of distortion proposed in the literature, will be discussed in the next section.

Ground-Truth Definition: A dataset for performance evaluation is not just a set of images gathered in a directory. Every image must be labeled with enough information to permit the comparison of the real output of an algorithm to the desired result for that image. This information is what is commonly called ground-truth. Therefore, the design of a dataset must also include the definition of which information is associated to every image and how this information will be represented. Ground-truth is something very specific to every application domain, for instance, bounding box coordinates for zone segmentation, a string of characters for text recognition, and the set of foreground pixels for binarization. In addition, not in every field it is absolutely clear which is the best way of defining and representing the ground-truth. In some areas, it has been a subject of intense debate among researchers. The different alternatives of ground-truth for each application domain will be presented and discussed in section "Public Databases for Document Analysis and Recognition." Some formats that have been proposed to represent the ground-truth will be described in section "Formats and Standards."

Ground-Truth Annotation: As pointed out before, every image of the dataset must be annotated with the ground-truth information. Doing this manually on a large set of images can be a very tedious and costly task and may be prohibitive if the ground-truth is complex. For that reason, dataset developers have used different strategies to reduce this cost. One way is obviously to use data generated synthetically. Then, the ground-truth can be obtained automatically. In real images, some strategies also exist, such as developing interactive and/or collaborative tools to help in the process of annotation or relying on the result of existing document analysis algorithms so that the user only has to confirm or correct the proposed ground-truth. These strategies will be discussed in more detail in the next section.

File Formats: Another important issue in the design of a dataset concerns the format of the files used to store images and represent the ground-truth. These formats should be well-known to the researchers and easy to manipulate to help in the dissemination and utilization of the dataset. In the case of input images, usual image formats, such as TIFF or JPEG, are commonly used, except in some particular applications as, for instance, online recognition. In the case of the ground-truth, there are more options, although XML has become a powerful pseudo-standard. Different specific formats proposed both for input images and ground-truth representation will be introduced in section "Formats and Standards."

Structure and Organization of the Database: Finally, all the elements related to a dataset (basically images and ground-truth) have to be put together, organized, and made available to the researchers in a way that permits an easy access and manipulation. Sometimes images are organized into different subsets according to several criteria, such as the category of the document, the degree of noise, or the difficulty of analysis. In any case, it is highly recommended to provide standard training and validation/test sets along with recommendations of use to permit a fair comparison of all algorithms executed on the dataset. These topics are very specific to each dataset and will be discussed in section "Public Databases for Document Analysis and Recognition."

Data Collection and Annotation

Real Data vs. Synthetic Data

There are basically two possibilities for collecting data: to use real data or to generate synthetic data.

Clearly, the main advantage of using real data is that it permits to evaluate the algorithms with the same kind of images encountered in real applications. Thus, evaluation can be a very good estimate of performance in real situations. However, manually collecting a large number of real images is a great effort that can be unaffordable in some cases. In addition, the annotation of these images with their corresponding ground-truth is also very costly in terms of time and human resources,

and errors can easily be introduced by manual annotation. Another disadvantage in some domains can be the difficulty of having easy access to a sufficient number of real images. Sometimes, confidentiality issues can make it difficult to provide public and free access to some collections of real data. Moreover, it is not easy to quantify the degree of noise in a real image. Then, it is not possible to define a ranking of difficulty of images according to the degree of noise.

The alternative to real data is to develop automatic methods to generate synthetic data. Clearly, the main advantages of this approach are that it allows to generate as many images as necessary and that the annotation of images with the ground-truth is also automatic. Then, manual effort is reduced. In addition, images generated using these methods can be easily classified according to the type and degree of noise or degradation applied, permitting to assess the reduction in performance with increasing degrees of image degradation. However, the main difficulty is to succeed in the development of models that could generate data as similar as possible to real data taking into account all kinds of noise and deformations. In some domains, this can be relatively easy, but in some other domains, this is really a challenging task.

Collection and Annotation of Real Data

The annotation of large sets of real data is always a costly process. In the literature one may find basically two kinds of approaches to reduce this effort. The first one still relies on using some interactive and/or collaborative tool to help in the process of annotation. It still requires a significant amount of human effort but tries to make it more comfortable and reduce the number of errors. The second approach tries to take advantage of existing techniques in document analysis. It consists in applying such existing methods to generate a first version of the ground-truth that later is revised and, if needed, corrected by a human user.

Tools for Manual Annotation of Data

Some of the first efforts in creating datasets for zone segmentation and text recognition relied heavily on collaborative human intervention [51, 58] for ground-truthing. This was a really costly process. For instance, a rough estimation of the time required to fully annotate 1 page was about a bit more than 1 h per page [51], only for zone segmentation. Moreover, manual annotation usually requires that more than one person creates and/or validates the ground-truth of a given page.

In order to reduce the amount of time required, several interactive tools have been developed and described in the literature. A significant group of such tools [3, 30, 77] deal with the definition and edition of zones in a document. In these tools zones can be represented in different ways such as rectangles, isothetic, or arbitrary polygons. They also permit to associate some kind of information to every zone, such as the type of contents, geometric and semantic attributes, or relation with other zones, for instance, reading order. All this information is stored using some specific format, usually based on XML. Most of these tools were originally conceived to annotate zone information for page segmentation evaluation. However, zones are

a common way to represent elements in document analysis. Therefore, a flexible representation of the zones and their associated information would permit to use these tools for other applications, such as logo detection or graphics recognition.

In contrast to zone-based annotation, many other applications (such as binarization or segmentation) require labeling at pixel level. PixLabeler [60] provides a tool to define such a ground-truth. Pixel labels (number and name) are totally configurable. Labeling of pixels can be done either manually or automatically using specific image-processing operations.

Graphics recognition is another domain where the complex layout and diversity of entities make annotation a costly process. In this context, a collaborative tool to annotate graphics documents has been developed in the framework of the project EPEIRES (available at http://www.epeires.org) for the tasks of symbol recognition and localization. Being a collaborative tool, users can contribute their own images, annotate documents, or validate the ground-truth labeled by other users.

For text recognition, the labeling is just the transcription of the text. It is difficult to develop tools to help in the process of transcription. Sometimes, an initial automatic transcription is used as a basis for correction (see next section). An alternative is to use predefined forms or templates that a selected group of users has to replicate. In this case, the ground-truth is implicit in the template, and the correspondence between the writing of the users and the template is usually easy to find. When this is not the case, for instance in online drawings [25], interactive tools can also be used.

Some interactive tools also exist to create ground-truth for table recognition [24, 65] that permit to easily mark row and column separators, indicate cells spanning various rows or columns, and represent the table as a hierarchical structure.

Semiautomatic Ground-Truthing

In this section we will describe several strategies that have been used in order to somehow automatize the process of ground-truth creation. These strategies depend very much on the application domain of the dataset, but, in general, they consist in applying either some baseline method that permits to automatically obtain a first version of the annotation that later is revised by a human user, either applying some pre-processing steps that allow to extract some components or information that makes the annotation easier. In some cases, they also imply some constraints or requirements in the acquisition of data.

One generic method to get automatically the ground-truth was proposed in [29] when images in the dataset are obtained by transforming somehow real ground-truthed images. In this case, images were generated by printing, and later, faxing, photocopying, and re-scanning original images with geometric ground-truth available at word and zone level. The ground-truth of the new generated images was obtained putting in correspondence a set of feature points in the original and the new generated images and finding the best affine transformation between the two sets of points. Feature points were the coordinates of the bounding boxes of groups of connected components obtained by joining nearby components.

Handwritten text recognition is a specific domain where semiautomatic ground-truthing makes special sense. In most cases, ground-truth simply consists of the labels of the characters or words. Many datasets have been constructed by asking a group of people to transcribe a predefined text. In these cases, a careful design of the acquisition conditions can permit an easy extraction of the text and the creation of the ground-truth by just aligning the input text with its expected predefined transcription. Only some basic image-processing operations, such as binarization, skew correction, or sometimes, simple character or line segmentation, are required [32, 75]. At the end of the process, a human user can verify the ground-truth to detect possible errors in the transcription or in the automatic extraction of data. A bit more complicated is the case where text can consist of several lines. Then, the alignment of the transcription with the input text has to take into account line breaks that are usually marked manually [39]. A similar approach is also used for annotating handwritten music scores [16] where the handwritten symbols are easily extracted after staff removal.

When the data is obtained from real scanned forms, there is no expected transcription to be matched with the input text. Then, a recognition engine trained on a different dataset can be applied to obtain a first version of the ground-truth to be manually validated and corrected [15, 28]. This strategy has also been used in other domains such as recognition of mathematical expressions [18] or chart recognition [76].

Binarization is another specific domain where it is not easy to define and create the ground-truth. In [44] the authors propose an alternative for getting the ground-truth that takes advantage of the intrinsic properties of text. The image is first skeletonized in order to obtain a silhouette of one pixel wide of the characters. The user checks and corrects the result of the skeletonization. Then, the skeleton is dilated in successive iterations until it reaches the edges of the character in the original image.

Generation of Synthetic Data

The generation of synthetic data to be used for performance evaluation of real systems depends on the existence of accurate models able to reproduce noise and variability of real data. Several models have been proposed in the literature. They can be broadly grouped in two main categories. A first group of models try to reproduce the distortion yielded by the process of acquisition of the image (printing, photocopying, scanning). A second group aims at creating models of the variability of the data. Variability can be due to several factors, such as handwriting, noise, or different spatial layout of objects in the image.

A first attempt to create a model of the distortion due to the acquisition process is presented in [6]. The model has several parameters modeling diverse factors that can influence the acquisition of an image such as the spatial sampling rate (digitizing resolution), blurring, digitizing threshold, sensitivity of pixel sensors,

skew, stretching, and translation. In a similar way, Kanungo proposed another framework for both global and local models of document degradation due to acquisition factors [27]. The global model focused on the degradation motivated by scanning thick books. The model takes into account the bending of the document that produces a perspective distortion and a change of intensity and blurring on the curved parts of the page. The local model accounts for pixel inversion (from foreground to background and vice versa), due to lighting changes, sensitivity of the sensors, and thresholding, and for blurring due to the scanner optical system. Apart from this theoretical models, simple image transformations (such as salt and pepper noise, rotation, blurring, erosion, dilation, slant shrink, or downsampling) have also been applied to generate noisy images in different domains such as line drawings [78] or music scores [10].

The second group of models is related to the generation of handwriting. We can distinguish different strategies to approaching this problem. Several works [52, 61] have explored the definition of explicit models of the human movement of the hand and the pen. These models permit to determine the trajectory of the pen when writing different letters or strokes. Alternatively, other works try to model, not the explicit trajectory, but the result of that trajectory that is, the shape of the stroke or the character. For instance, B-splines can be used [74] to define a generative model of each letter (decomposed into three parts, the head, the body, and the tail) where the distribution of the control points of the B-splines is learned from a set of existing samples. A deformable model is used to smoothly join the tail and the head of two adjacent letters.

In a completely different approach, other works do not try to define a model of handwriting but just to generate new samples similar to other existing ones. This can be done in different ways. One possibility for generating individual characters is using point correspondence among several samples [42]. A template of the character is created as the mean of all the samples. Then, a point correspondence can be established between every sample and this template. New samples can be generated by moving original points along the line that joins every point in the sample with the corresponding point in the template. Distortion can also be generated at the line level [71]. In this case four kinds of pixel level transformations (shearing, bending, and horizontal and vertical scaling) are applied globally to the whole line of text. The amount of transformation applied to every pixel depends on its horizontal position and is determined by a set of underlying functions based on the cosine. Additionally, horizontal and vertical scaling is also applied to the connected component level. Finally, thinning/thickening effects are applied to the whole line. At the word level for online handwriting, the work presented in [63] finds patterns between velocity of writing and the shape of corners, the slant and the size. Distortions are generated by changing the velocity pattern from an original image of the word. There is a limit in the amount of distortion based on keeping some features of the word, such as ascenders and descenders.

A last strategy [23] that has been used for the generation of handwriting consists in the concatenation of models of characters or groups of characters. The basic idea

is decomposing a word into a number of basic units (common groups of characters) trying to minimize the number of cuts. New samples of the word are generated by concatenating existing samples of the characters trying to optimize the transition between groups.

In other domains, specific methods have also been proposed with the goal of generating synthetic images as similar as possible to real ones. For instance, in the case of binarization [47], synthetic images are generated by overlapping a clean image with noisy ones. These noisy images correspond to the scanning of blank pages from the eighteenth century that contain usual artifacts such as stains, strains, or uneven illumination.

In addition to noise and degradation at the pixel level, another important source of variability in document images is that of the spatial layout of entities in the document. However, not a lot of attention has been devoted to this matter in the generation of synthetic data. Grammars have been used [38] to generate sketched mathematical expressions. In the field of graphics recognition, a rule-based system [11] has been proposed to generate graphic documents (architectural floor plans and electronic diagrams) based on overlapping entities (in this case, symbols) on a set of predetermined, fixed backgrounds, following a set of domain-specific rules that define where and how symbols can be placed.

Formats and Standards

XML has become a de facto standard to represent ground-truth data in many existing datasets. Many datasets use a specific XML schema for ground-truth in several domains. In some cases some specific standard formats derived from XML have been defined and used to represent ground-truth as it is the case of ALTO (Analyzed Layout and Text Object, at http://www.loc.gov/standards/alto/) or PAGE [53] for page analysis, InkML for online text recognition [25], MathML for representing mathematical expressions [45], or SVG (Scalable Vector Graphics, at http://www.w3.org/Graphics/SVG/) for graphics recognition.

However, some efforts have also been devoted to creating specific formats in some applications, such as DAFS [17] and RDIFF [77] for page analysis, UNIPEN [21] for online recognition, and VEC for graphics recognition [49].

Public Databases for Document Analysis and Recognition

This section intends to provide a comprehensive summary of available public datasets for document analysis tasks, grouped in five categories: document imaging (mainly binarization and dewarping), page analysis (basically, page/zone segmentation, and document classification), text recognition (both printed and handwritten, offline and online), graphics recognition (including vectorization,

symbol recognition and spotting, music scores, chart and sketch recognition), and other applications (table recognition and understanding, mathematical expressions, signature verification). For each of these categories, several generic aspects concerning the generation of datasets for that domain are discussed, such as collection and acquisition of data and ways of defining the ground-truth for that particular domain. A table summarizing the public datasets for each category is included with information such as the type of data included in the dataset, the size, training, and test sets. It is worth noticing that some of these datasets are also referred, and sometimes further described, throughout the book in the chapters devoted to each particular subfield of document analysis. In some chapters (see for instance, ▸Chap. 7 (Page Similarity and Classification)), other non-public datasets are also mentioned to show the state-of-the-art in a particular field. In this section, however, we will only include public datasets. In a notes section at the end of this chapter, we also provide the links to the websites hosting the datasets. Reference to papers is only included when they cannot be found in an institutional website.

Document Imaging

Binarization

Two main issues arise concerning the creation and annotation of datasets for binarization: the definition of the ground-truth and the collection/generation of the data.

The definition of the ground-truth is closely related to the kind of evaluation approach. In [44] different approaches to the evaluation of binarization algorithms are distinguished. The first category is a human-centered evaluation where binarization is evaluated by visual inspection of the result according to a set of predefined criteria. Therefore, there is no need for ground-truth. However, the evaluation is subjective and time consuming and lacks robustness. The second category is the so-called goal-directed evaluation where the result of binarization is evaluated through its impact in the performance of a high-level module, usually OCR. In this case, ground-truth must be defined in terms of the output of this high-level module (i.e., usually at character level). In the third category the evaluation is done directly on the result of the binarization at pixel level, and, therefore, ground-truth must also be defined at pixel level. The definition of the ground-truth at character level is usually easier and more objective than defining it at pixel level. It also avoids having to define a set of subjective criteria for evaluating the performance. However, the evaluation could depend on the specific method selected for implementing the high-level module. The definition of the ground-truth at pixel level is the only way to evaluate the output of the binarization itself, and this is how ground-truth is defined in all public datasets described in this section. However, it is not easy to define this kind of ground-truth since it is very subjective. A study analyzing the influence of different ways of ground-truthing in performance evaluation of binarization can be found in [66].

Table 29.1 Binarization datasets

Name	No. images	Type of images
DIBCO'09 [19][1]	10	Real images
H-DIBCO'10 [54][2]	10	Real images
DIBCO'11 [55][3]	16	Real images
ICFHR'10 Contest [47][4]	60 training/210 test	Synthetic images

Table 29.2 Image dewarping datasets

Name	No. images	Type of images
Document Image Dewarping Contest [64][5]	102	Real images

The collection of large amounts of real data is not easy, especially when ground-truthing has to be done at pixel level, since this is a time-consuming process. Therefore, existing datasets rely mainly on using synthetic images [47] or on semiautomatic methods [44] for the generation of ground-truth.

Table 29.1 summarizes existing datasets for evaluation of binarization methods. They have been generated for their use in competitions organized in the framework of international conferences, such as ICDAR (in the case of the series of DIBCO datasets) and ICFHR. In the DIBCO dataset, real images (including machine-printed and handwritten text) are used, obtained from collections of different libraries, containing a set of representative degradations such as variable background intensity, shadows, smear, smudge, low contrast, bleed-through, or show-through. The ICFHR 2010 dataset consists of synthetic images generated by the combination of clean images with noisy ones.

Dewarping

For page dewarping, one public dataset exists, created in the context of the Document Image Dewarping Contest organized at CBDAR'07 [64]. It is composed of real images from several technical books captured by an off-the-shelf digital camera in a normal office environment, binarized using a local adaptative thresholding technique. Three different kinds of ground-truth are provided. Two of them (text lines and text zones) are intended to serve as internal evaluation of the methods as most of them perform an intermediate step consisting in the detection of curved text lines. The third kind of ground-truth information is the ASCII transcription of the text in the document that can be used to perform a global goal-oriented evaluation based on OCR accuracy. Details are summarized in Table 29.2.

Page Analysis

The evaluation of page analysis algorithms (mainly physical layout analysis or zone segmentation) has received a lot of attention throughout the years. An important issue that has been largely discussed in many works is the type of ground-truth

information required, as mentioned in ▸Chap. 5 (Page Segmentation Techniques in Document Analysis). The answer to this question has influence in many aspects of the evaluation framework, including the goal of the evaluation process, the generation of the dataset, or the evaluation measures. Three alternatives are described in the literature: text-based, pixel-based, or zone-based ground-truth information.

Text-based evaluation uses the transcription of the document as a ground-truth. The main advantage of this approach is simplicity as the labeling information is easy to obtain and store, and there is no need for a specific output format of zoning systems. Although it was used in some of the early works on zoning evaluation [26], it was later dismissed due to multiple drawbacks. It is not a direct evaluation method for page segmentation results, and thus, it does not permit to provide any characterization of the segmentation errors. In addition, it can only deal with text regions and requires the zoning algorithm to be part of an OCR system.

Pixel-based approaches [60] define the ground-truth in terms of pixel labels. Every pixel is assigned a different label depending on the zone they belong to. In this way, they do not depend on an accurate definition of the shape of the zones and can handle better noise or graphics elements. They are more appropriate to evaluate zone classification (label assigned to every pixel) than zone segmentation (segmentation of the document into zones).

Zone based is the most used approach for representing ground-truth information. The ground-truth is defined in terms of the set of physical zones that compose the document. The shape of the zones has been described mostly using rectangles [30], but also isothetic polygons [4] or arbitrary polygons [69]. In this way the ground-truth contains all the information needed to evaluate zone segmentation as a stand-alone process and characterize the output of the system according to different segmentation errors. In addition, a label can be added to every zone describing its contents and allowing for evaluation of zone classification.

In any case, one of the critical issues is the criteria followed to define the zones or the regions in the ground-truth. In some cases, a set of rules is defined to limit the subjectivity of ground-truthers when determining the zones of a document and their shape [41, 69].

XML has been mostly used to define the format to store the ground-truth files, although other specific formats have also been defined such as DAFS [31] or RDIFF [77]. In some cases there is also an internal graph structure [30, 69] that permits to represent the logical order of the zones.

Table 29.3 summarizes the details of existing public datasets in terms of the type of documents included in the datasets, ground-truth information, and tasks for which the dataset could be used (segmentation, page classification, or logical layout analysis). All datasets contain real images. Only one dataset, the UVA dataset [69], includes color images. A slightly different dataset is that released for the Book Structure Extraction Competition held at ICDAR'09 and ICDAR'11 [12]. This dataset is intended to evaluate the detection of the logical book structure in terms of the table of contents.

Table 29.3 Page analysis datasets

Name	No. images	Type	Ground-truth	Task
UW-III	1,600	Technical	Rectangle	Segmentation/OCR
ICDAR'09[6]	1,240	Magazines/ journals	Isothetic pols.	Segmentation
ICDAR'11 [5][7]	100	Historical	Isothetic pols.	Segmentation
MARG[8]	3,337	Journals	Rectangle	Segmentation/ classification
UvA [69][9]	1,000	Color magazines	Arbitrary shapes	Segmentation/ logical order
Book structure [12][10]	1,040	Books	ToC	Logical structure

Text Recognition

Text recognition has traditionally been the main activity within the field of document analysis. Therefore, there has been a lot of work around the creation of datasets for evaluation of text recognition methods. Tables 29.4 and 29.5 compile a comprehensive list of all these existing datasets for text recognition. They are organized according to the writing script since usually, methods for recognition follow different approaches depending on the particularities of each script (see ▸Chaps. 9 (Language, Script, and Font Recognition), ▸13 (Middle Eastern Character Recognition), and ▸14 (Asian Character Recognition)). The table also includes the type of acquisition of images (online/offline) and a summary of the contents and the number of samples contained in each dataset. For more details, we refer the reader to the original paper describing the dataset or the dataset website. Some of these datasets have also been described in the chapters devoted to text recognition (see ▸Chaps. 12 (Continuous Handwritten Script Recognition), ▸13 (Middle Eastern Character Recognition), and ▸14 (Asian Character Recognition)). It is worth mentioning that several of these datasets have been used in international competitions or evaluation campaigns.

Additionally, in Table 29.6 we list two existing datasets for evaluation of writer identification for Arabic and Latin text. Both datasets have been used in competitions organized in the framework of last ICDAR conference.

Finally, and also related to text recognition, we present in Table 29.7 one dataset that has been used to evaluate word and line segmentation in ICDAR competitions.

Graphics Recognition

The field of graphics recognition is very diverse, covering different topics such as vectorization, symbol recognition and spotting, drawing interpretation, and music score analysis. Therefore, the existing datasets also reflect this diversity. The first efforts in providing standard datasets for graphics recognition were related to vectorization and arc detection and segmentation, in the context of the GREC workshop.

Table 29.4 Text recognition datasets (1)

Name	Script	Mode	Contents	Volume
UW-III dataset [51][11]	Latin	Offline	Printed text	≈1,000 pages
Noisy text [35][12]	Latin	Offline	Noisy printed text	3,305 pages
ISRI-UNLV [58][13]	Latin	Offline	Printed characters + OCR-based zone segmentation	≈2,000 pages
Sofia-Munich corpus [40]	Latin Cyrillic	Offline	Printed characters	2,306 pages
CENPARMI database [68]	Latin	Offline	HW digits	≈17K digits
CEDAR database[14]	Latin	Offline	HW characters, zip codes, digits, city names, state names	≈1,000 words, 10K codes, 50K characters
NIST database[15]	Latin	Offline	HW digits, characters, sentences	≈810K characters
MNIST database[16]	Latin	Offline	HW digits	70K digits
IAM-database[17]	Latin	Offline	HW words/lines/sentences	82,227 words/9,285 lines
RIMES dataset[18]	Latin	Offline	HW words and lines	≈250K words
IAM-HistDB[19]	Latin	Offline	HW words/sentences – historical documents	6,543 lines/39,969 words
George Washington dataset [57][20]	Latin	Offline	HW words for word spotting	20 pages/5,751 words
The Rodrigo database [62][21]	Latin	Offline	HW lines – historical documents	≈20K lines/231K words
The Germana database [48][22]	Latin	Offline	HW lines – historical documents	≈21K lines/217K words
IRONOFF database [72][23]	Latin	Offline Online	HW characters, digits, words	≈32K characters/50K words
UNIPEN project [21][24]	Latin	Online	HW characters, digits, words, text	≈16K digits, 300K characters, 160K words, 16K sentences
IAM-OnDB[25]	Latin	Online	HW sentences	≈86K words/13K lines
IAM-OnDo[26]	Latin	Online	HW varied documents: text, tables, formulas, diagrams, drawings, markings	1,000 documents

Symbol recognition has been another active field in proposing datasets for performance evaluation. Going beyond symbol recognition, symbol spotting appears as a challenging task for the graphics recognition community. Two datasets have been released in the last years that can be used for the evaluation of symbol spotting,

Table 29.5 Text recognition datasets (2)

Name	Script	Mode	Contents	Volume
APTI database[27]	Arabic	Offline	Printed words	45,313,600 words
IFN/ENIT[28]	Arabic	Offline	HW words	≈26,400 city names
AHDB database[29]	Arabic	Offline	HW words and sentences	
IBN-SINA dataset[30]	Arabic	Offline	HW words – word spotting	≈1,000 subwords
ADAB database [13]	Arabic	Online	HW words	15,158 words
CENPARMI dataset [2]	Arabic	Offline	HW digits and words (from cheques)	23,325 words/9,865 digits
CENPARMI dataset [1]	Arabic	Offline	HW digits, numeral strings, isolated letters, words, dates	46,800 digits, 13,439 strings, 21,426 letters, 11,375 words, 284 dates
Hoda database[31]	Farsi	Offline	HW digits	102,352 digits
CENPARMI dataset [67]	Farsi	Offline	HW digits, numerical strings, legal amounts, letters, and dates	18,000 digits, 7,350 strings, 11,900 letters, 8,575 words, 175 dates
IFHCDB database[32]	Farsi	Offline	HW characters and numerals	52,380 characters/17,740 numerals
HIT-MW dataset[33]	Chinese	Offline	HW characters, documents	186,444 characters
SCUT-COUCH2009 dataset[34]	Chinese	Online	HW characters words	1,392,900 characters
CASIA dataset[35]	Chinese	Online Offline	HW characters, documents	≈3.9 million characters/ 5,090 pages
[7][36]	Indian scripts	Online Offline	HW numerals and characters	45,948 numerals

Table 29.6 Writer identification datasets

Name	Alphabet	Volume
Arabic WI contest [22][37]	Arabic	54 writers/162 paragraphs
ICDAR'11 WI contest [36][38]	Latin	26 writers/208 pages

Table 29.7 Word and line segmentation datasets

Name	Alphabet	Contents	Volume
ICDAR'09 HW contest [20][39]	Latin	Lines and words	300 images

the first one based on images of scanned real drawings and a second one based on a semi-automatic method of generation of drawings that has allowed the generation of an extense dataset composed of synthetic images of architectural floor plans and electronic diagrams.

In the field of music interpretation, datasets exist for two tasks, namely, staff removal and writer identification in handwritten music scores. Finally, datasets also exist in the context of chart analysis, to evaluate the process at different levels:

Table 29.8 Datasets for graphics recognition

Name	Domain	Tasks	Contents
Arc segmentation contests[40]	Engineering drawings	Arc segmentation	
SymbolRec[41]	Symbols	Symbol recognition	7,500 images of 150 symbols
NicIcon[42]	Symbols	Symbol recognition	Online/offline. 14 classes. 26,163 images
FPLAN-POLY [59][43]	Architectural drawings	Symbol spotting	42 floor plan images + 38 query symbols
SESYD [11][44]	Architectural and electronic drawings	Symbol spotting	2,000 images/≈40,000 symbols
CVC-MUSCIMA[45]	Music scores	Writer identification staff removal	1,000 scanned images (50 writers × 20 documents) + 12,000 synthetic images
Synthetic Score Database[46]	Music scores	Staff removal	Synthetic images from 32 generated music scores
CHIME Chart Image [76][47]	Charts	Chart interpretation	200 real images + 3,200 synthetic images

vector level (lines and arcs), text level (blocks and words), and chart level (axis, titles, labels, data points).

Table 29.8 summarizes the main features of all these datasets.

Other Applications

In this section we describe datasets for miscellaneous document analysis applications that did not fit in any of the previous sections.

Firstly, we encounter a set of datasets for text localization and recognition in images obtained from non-paper sources, such as web images or scene images acquired with a camera. Some of these datasets have been used in the series of competitions on robust reading in several editions of ICDAR starting in 2003. In addition, several other datasets have been created in the last years, mainly due to the rise of digital cameras that have led to the existence of a large amount of natural images containing text. The characteristics of these datasets are summarized in Table 29.9 and also referred to in ▸Chap. 25 (Text Localization and Recognition in Images and Video).

In a second group there is a set of datasets for recognition of mathematical expressions, both symbols and relations among them, to derive the structure of the expression. Table 29.10 summarizes these datasets.

Another important group concerns datasets for signature verification, mostly used in competitions organized in the context of ICDAR and ICFHR. Two important issues to take into account in this domain (see ▸Chap. 27 (Online Signature Verification))

Table 29.9 Datasets for robust reading

Name	Type of images	Tasks	Contents
ICDAR'03 and ICDAR'05 [37][48]	Scene images	Text localization, character recognition, and word recognition	
ICDAR'11 Challenge 1[49]	Web and e-mail images	Text localization, segmentation and word recognition	420 images (3,583 words) for training and 102 (918 words) for test
ICDAR'11 Challenge 2[50]	Scene images	Text localization and word recognition	485 images and 1,564 words
KAIST database[51]	Scene images	Text localization and segmentation	3,000 images
Google Street View dataset [73][52]	Scene images	Word spotting	350 images
Street View House Numbers Dataset[53]	Digits from natural images	Digit recognition	600,000 digits
IUPR dataset [9][54]	Document pages	Line/zone segmentation, dewarping, text recognition	100 images
NEOCR dataset[55]	Scene images	Text localization and recognition	659 images
Chars74K dataset[56]	Characters from natural images	Character recognition	74,000 characters

Table 29.10 Datasets for recognition of mathematical expressions

Name	Type of images	Contents
Infty [70][57]	Printed text	108,914 words and 21,056 mathematical expressions. 688,580 character samples, from 476 pages of text
[18]	Printed text	400 images (297 real/103 synthetic). 2,459 displayed and 3,101 embedded expressions
MathBrush[58]	Offline handwritten	4,655 expressions
CROHME [43][59]	Online	921 expressions for training and 348 for test
UW-III dataset [51][60]	Offline	100 expressions from 25 pages
Hamex [56][61]	Online	4,350 expressions
Marmot[62]	PDF documents	9,482 expressions

are the collection of a representative set of genuine and forgery signatures and the specific protocol for using reference, genuine and forgery signatures. Table 29.11 summarizes the datasets for signature verification. Some of them are further described in ▸Chap. 27 (Online Signature Verification).

Finally, the last group contains a pair of datasets for table processing, mainly table localization and cell segmentation. The first one uses the UNLV and UW-III datasets that have been described in section "Text Recognition." The ground-truth for table processing has been generated and released[70] for 427 images in the UNLV dataset and 120 images in the UW-III dataset [65]. The second one has been

Table 29.11 Datasets for signature verification

Name	Type	Domain	Writers	Genuine	Forgeries
GPDS-960[63]	Offline	Western	960	24 per writer (23,049)	30 per writer (28,800)
MCYT [46][64]	Offline Online	Western	330	5 per writer (1,650)	25 per writer (8,250)
BioSecurId [14][65]	Offline Online	Western	400	16 per writer (6,400)	12 per writer (4,800)
SVC2004[66]	Online	Western Chinese	100	20 per writer (2,000)	20 per writer (2,000)
SigComp09 [8][67]	Offline Online	Western	12 (training) 100 (test)	60 (training)/ 1,200 (test)	1,860 (training)/ 753 (test)
4NSigComp10 [33][68]	Offline	Western	1	85 (training)/ 28 (test)	124 (training)/ 97 (test)
SigComp11 [34][69]	Offline Online	Western Chinese	10 (Chinese) 54 (Western)	240 (Chinese)/ 1,330 (Western)	400 (Chinese)/ 630 (Western)
Caltech	Camera based	Western	105	25–30 per writer	10 per writer

generated for an ICDAR'11 competition using PDF documents (19 for training, 3 for validation and 19 for test) containing in total 1,003 tables for training, 120 for validation and 968 for test.[71]

Conclusion

The need of standard and public tools for a fair evaluation of all the existing methods in document analysis has become an evidence in the last years. The creation of representative datasets is the first step in this process. Thus, the number of such datasets has largely increased along the last years in all fields of document analysis. Currently, we can find at least one dataset to evaluate almost any document analysis algorithm. This will certainly help to boost the research results as any contribution can be easily put in the context of the current state-of-the-art.

However, the creation of datasets arises important issues that, in some cases or domains, are far from to be completely solved. Throughout this chapter we have discussed the more important ones, mainly concerned with the collection of a representative and sufficiently large number of images and with the annotation of such large data. We have analyzed the different approaches used for that, and which have allowed to generate really large annotated datasets.

As a result of all this work, many datasets have been generated in all fields of document analysis. We have summarized all the existing public datasets in an organized way, according to the domain and we have described the main characteristics of each dataset. We can see this large number of datasets as a consequence of mainly two factors: on one hand, the own evolution of each field, demanding every time for more complex and complete datasets to evaluate new advances in the field, and, on

the other hand, the diversity of each particular domain and, thus, the need of having datasets that cover all this diversity (types of documents, fonts, handwriting styles, layouts, noise, distortion, etc.). In this way, different aspects of the same problem can be evaluated, and the more challenging data can be identified. In this sense, probably, in many fields there is no need for more datasets but an identification of the most relevant and challenging subsets of data to focus new research on it.

Cross-References

▸A Brief History of Documents and Writing Systems
▸Asian Character Recognition
▸Continuous Handwritten Script Recognition
▸Middle Eastern Character Recognition
▸Online Signature Verification
▸Page Segmentation Techniques in Document Analysis
▸Page Similarity and Classification
▸Text Localization and Recognition in Images and Video
▸Tools and Metrics for Document Analysis Systems Evaluation

Notes

Links to datasets for binarization

[1]DIBCO'09: http://www.iit.demokritos.gr/bgat/DIBCO2009/benchmark

[2]H-DIBCO'10: http://www.iit.demokritos.gr/bgat/H-DIBCO2010/benchmark

[3]DIBCO'11: http://utopia.duth.gr/ipratika/DIBCO2011/benchmark

[4]ICFHR'10 Contest: http://users.dsic.upv.es/iaprtc5/icfhr2010contest/index.php

Links to datasets for image dewarping

[5]Document Image Dewarping Contest: http://www.dfki.uni-kl.de/shafait/downloads.html

Links to datasets for page analysis

[6]ICDAR'09: http://dataset.primaresearch.org/

[7]ICDAR'11: http://www.prima.cse.salford.ac.uk:8080/ICDAR2011_competition/

[8]MARG: http://marg.nlm.nih.gov/index2.asp

[9]UvA: http://www.science.uva.nl/UvA-CDD/

[10]Book structure: http://users.info.unicaen.fr/doucet/StructureExtraction/2011/

Links to datasets for text recognition

[11]UW-III dataset: available on CD-ROM

[12]Noisy Text: http://www.cse.lehigh.edu/lopresti/noisytext.html

[13]ISRI-UNLV: http://code.google.com/p/isri-ocr-evaluation-tools/

[14]CEDAR database: http://www.cedar.buffalo.edu/Databases/index.html

[15]NIST database: http://www.nist.gov/srd/nistsd19.cfm

[16]MNIST database: http://yann.lecun.com/exdb/mnist/

[17]IAM-database: http://www.iam.unibe.ch/fki/databases/iam-handwriting-database

[18]RIMES dataset: http://www.rimes-database.fr/wiki/doku.php
[19]IAM-HistDB: http://www.iam.unibe.ch/fki/databases/iam-historical-document-database
[20]George Washington dataset: http://ciir.cs.umass.edu/cgi-bin/downloads/downloads.cgi
[21]The Rodrigo database: http://prhlt.iti.upv.es/page/projects/multimodal/idoc/rodrigo
[22]The Germana database: http://prhlt.iti.es/page/projects/multimodal/idoc/germana
[23]IRONOFF database: http://www.irccyn.ec-nantes.fr/viardgau/IRONOFF/IRONOFF.html
[24]UNIPEN project: http://www.unipen.org/products.html
[25]IAM-OnDB: http://www.iam.unibe.ch/fki/databases/iam-on-line-handwriting-database
[26]IAM-OnDo: http://www.iam.unibe.ch/fki/databases/iam-online-document-database
[27]APTI database: http://diuf.unifr.ch/diva/APTI/
[28]IFN/ENIT: http://www.ifnenit.com/index.htm
[29]AHDB database: http://www.cs.nott.ac.uk/cah/Databases.htm
[30]IBN-SINA dataset: http://www.causality.inf.ethz.ch/al_data/IBN_SINA.html
[31]http://farsiocr.ir/farsi-digit-dataset
[32]IFHCDB database: http://ele.aut.ac.ir/imageproc/downloads/IFHCDB.htm
[33]HIT-MW dataset: http://sites.google.com/site/hitmwdb/
[34]SCUT-COUCH2009 dataset: http://www.hcii-lab.net/data/scutcouch/EN/introduction.html
[35]CASIA dataset: http://www.nlpr.ia.ac.cn/databases/handwriting/Home.html
[36]http://www.isical.ac.in/ujjwal/download/database.html

Links to datasets for writer identification

[37]Arabic WI contest: http://wic2011.qu.edu.qa/
[38]ICDAR'11 WI contest: http://users.iit.demokritos.gr/louloud/Writer_Identification_Contest/

Links to datasets for word and line segmentation

[39]ICDAR'09 HW contest: http://users.iit.demokritos.gr/bgat/HandSegmCont2009/Benchmark/

Links to datasets for graphics recognition

[40]Arc segmentation contests: http://www.cs.usm.my/arcseg2011/, http://www.cs.usm.my/arcseg2009/, http://www.cs.cityu.edu.hk/liuwy/ArcContest/ArcSegContest.htm
[41]SymbolRec: http://iapr-tc10.univ-lr.fr/index.php/symbol-contest-2011/153
[42]NicIcon: http://www.unipen.org/products.html
[43]FPLAN-POLY: http://www.cvc.uab.cat/marcal/FPLAN-POLY/index.html
[44]SESYD: http://mathieu.delalandre.free.fr/projects/sesyd/
[45]CVC-MUSCIMA: http://www.cvc.uab.es/cvcmuscima
[46]Synthetic Score Database: http://music-staves.sourceforge.net
[47]CHIME Chart Image: http://www.comp.nus.edu.sg/tancl/ChartImageDataset.htm

Links to datasets for robust reading

[48]ICDAR'03 and ICDAR'05: http://algoval.essex.ac.uk/icdar/Competitions.html
[49]ICDAR'11 Challenge 1: http://www.cvc.uab.es/icdar2011competition/
[50]ICDAR'11 Challenge 2: http://robustreading.opendfki.de/wiki/SceneText
[51]KAIST database: http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database
[52]Google Street View dataset: http://vision.ucsd.edu/kai/svt/
[53]Street View House Numbers Dataset: http://ufldl.stanford.edu/housenumbers/
[54]IUPR dataset: http://www.sites.google.com/a/iupr.com/bukhari/
[55]NEOCR dataset: http://www6.informatik.uni-erlangen.de/research/projects/pixtract/neocr/
[56]Chars74K dataset: http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/

Links to datasets for recognition of mathematical expressions

[57]Infty: http://www.inftyproject.org/en/database.html

[58]MathBrush: https://www.scg.uwaterloo.ca/mathbrush/corpus

[59]CROHME: http://www.isical.ac.in/crohme2011/index.html

[60]UW-III dataset: available on CD-ROM

[61]Hamex: http://www.projet-depart.org/

[62]Marmot: http://www.founderrd.com/marmot_data.htm

Links to datasets for signature verification

[63]GPDS-960: http://www.gpds.ulpgc.es/download/index.htm

[64]MCYT: http://atvs.ii.uam.es/databases.jsp

[65]BioSecurId: http://atvs.ii.uam.es/databases.jsp

[66]SVC2004:

[67]SigComp09: http://sigcomp09.arsforensica.org/

[68]4NSigComp10: http://www.isical.ac.in/icfhr2010/CallforParticipation4NSigComp2010.html

[69]SigComp11: http://forensic.to/webhome/afha/SigComp.html

Links to datasets for table processing

[70]http://www.dfki.uni-kl.de/shahab/t-truth/

[71]http://www.liaad.up.pt/acs/Competition.htm

References

1. Alamri H, Sadri J, Suen CY, Nobile N (2008) A novel comprehensive database for Arabic offline handwriting recognition. In: Proceedings of the 11th international conference on frontiers in handwriting recognition (ICFHR 2008), Montréal, pp 664–669
2. Al-Ohali Y, Cheriet M, Suen C (2003) Databases for recognition of handwritten arabic cheques. Pattern Recognit 36(1):111–121. doi:10.1016/S0031-3203(02)00064-X, URL: http://www.sciencedirect.com/science/article/pii/S003132030200064X
3. Antonacopoulos A, Karatzas D, Bridson D (2006) Ground truth for layout analysis performance evaluation. In: Proceedings of the 7th IAPR workshop on document analysis systems (DAS2006), Nelson. Springer, pp 302–311
4. Antonacopoulos A, Bridson D, Papadopoulos C, Pletschacher S (2009) A realistic dataset for performance evaluation of document layout analysis. In: 10th international conference on document analysis and recognition (ICDAR'09), Barcelona, 2009, pp 296–300. doi:10.1109/ICDAR.2009.271
5. Antonacopoulos A, Clausner C, Papadopoulos C, Pletschacher S (2011) Historical document layout analysis competition. In: 11th international conference on document analysis and recognition (ICDAR'11), Beijing, 2011
6. Baird HS (1995) Document image defect models. In: O'Gorman L, Kasturi R (eds) Document image analysis. IEEE Computer Society, Los Alamitos, pp 315–325. URL: http://dl.acm.org/citation.cfm?id=201573.201660
7. Bhattacharya U, Chaudhuri B (2009) Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. IEEE Trans Pattern Anal Mach Intell 31(3):444–457. doi:10.1109/TPAMI.2008.88
8. Blankers V, Heuvel C, Franke K, Vuurpijl L (2009) ICDAR 2009 signature verification competition. In: 10th international conference on document analysis and recognition (ICDAR'09), Barcelona, 2009, pp 1403–1407. doi:10.1109/ICDAR.2009.216

9. Bukhari SS, Shafait F, Breuel TM (2012) The IUPR dataset of camera-captured document images. In: Proceedings of the 4th international conference on camera-based document analysis and recognition (CBDAR'11), Beijing. Springer, Berlin/Heidelberg, pp 164–171
10. Dalitz C, Droettboom M, Pranzas B, Fujinaga I (2008) A comparative study of staff removal algorithms. IEEE Trans Pattern Anal Mach Intell 30:753–766. doi:http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.70749
11. Delalandre M, Valveny E, Pridmore T, Karatzas D (2010) Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems. Int J Doc Anal Recognit 13:187–207. doi:http://dx.doi.org/10.1007/s10032-010-0120-x, URL: http://dx.doi.org/10.1007/s10032-010-0120-x
12. Doucet A, Kazai G, Dresevic B, Uzelac A, Radakovic B, Todic N (2011) Setting up a competition framework for the evaluation of structure extraction from OCR-ed books. Int J Doc Anal Recognit 14:45–52. doi:http://dx.doi.org/10.1007/s10032-010-0127-3, URL: http://dx.doi.org/10.1007/s10032-010-0127-3
13. El Abed H, Kherallah M, Märgner V, Alimi AM (2011) On-line Arabic handwriting recognition competition: ADAB database and participating systems. Int J Doc Anal Recognit 14:15–23. doi:http://dx.doi.org/10.1007/s10032-010-0124-6, URL: http://dx.doi.org/10.1007/s10032-010-0124-6
14. Fierrez J, Galbally J, Ortega-Garcia J, Freire M, Alonso-Fernandez F, Ramos D, Toledano D, Gonzalez-Rodriguez J, Siguenza J, Garrido-Salas J, Anguiano E, Gonzalez-de Rivera G, Ribalda R, Faundez-Zanuy M, Ortega J, Cardeñoso-Payo V, Viloria A, Vivaracho C, Moro Q, Igarza J, Sanchez J, Hernaez I, Orrite-Uruñuela C, Martinez-Contreras F, Gracia-Roche J (2010) BiosecurID: a multimodal biometric database. Pattern Anal Appl 13:235–246. doi:10.1007/s10044-009-0151-4, URL: http://dx.doi.org/10.1007/s10044-009-0151-4
15. Fischer A, Indermühle E, Bunke H, Viehhauser G, Stolz M (2010) Ground truth creation for handwriting recognition in historical documents. In: Proceedings of the 9th IAPR international workshop on document analysis systems (DAS'10), Boston. ACM, New York, pp 3–10. doi:http://doi.acm.org/10.1145/1815330.1815331, URL: http://doi.acm.org/10.1145/1815330.1815331
16. Fornés A, Dutta A, Gordo A, Lladós J (2012) CVC-MUSCIMA: a ground truth of handwritten music score images for writer identification and staff removal. Int J Doc Anal Recognit 15(3), 243–251. doi:10.1007/s10032-011-0168-2, URL: http://dx.doi.org/10.1007/s10032-011-0168-2
17. Fruchterman T (1995) DAFS: a standard for document and image understanding. In: Proceedings of the symposium on document image understanding technology, Bowes, pp 94–100
18. Garain U, Chaudhuri B (2005) A corpus for OCR research on mathematical expressions. Int J Doc Anal Recognit 7:241–259. doi:10.1007/s10032-004-0140-5, URL: http://dl.acm.org/citation.cfm?id=1102243.1102246
19. Gatos B, Ntirogiannis K, Pratikakis I (2009) ICDAR2009 document image binarization contest (DIBCO 2009). In: 10th international conference on document analysis and recognition (ICDAR'09), Barcelona, 2009, pp 1375–1382. doi:10.1109/ICDAR.2009.246
20. Gatos B, Stamatopoulos N, Louloudis G (2011) ICDAR2009 handwriting segmentation contest. Int J Doc Anal Recognit 14:25–33. doi:10.1007/s10032-010-0122-8, URL: http://dx.doi.org/10.1007/s10032-010-0122-8
21. Guyon I, Schomaker L, Plamondon R, Liberman M, Janet S (1994) Unipen project of on-line data exchange and recognizer benchmarks. In: Proceedings of the international conference on pattern recognition, Jerusalem, pp 29–33
22. Hassaï andne A, Al-Maadeed S, Alja'am JM, Jaoua A, Bouridane A (2011) The ICDAR2011 Arabic writer identification contest. In: International conference on document analysis and recognition (ICDAR), Beijing, 2011, pp 1470–1474. doi:10.1109/ICDAR.2011.292
23. Helmers M, Bunke H (2003) Generation and use of synthetic training data in cursive handwriting recognition. In: Perales F, Campilho A, de la Blanca N, Sanfeliu A (eds) Pattern recognition and image analysis. Lecture notes in computer science, vol 2652. Springer, Berlin/Heidelberg, pp 336–345

24. Hu J, Kashi RS, Lopresti DP, Wilfong GT (2002) Evaluating the performance of table processing algorithms. Int J Doc Anal Recognit 4(3):140–153
25. Indermühle E, Liwicki M, Bunke H (2010) IAMonDo-database: an online handwritten document database with non-uniform contents. In: Proceedings of the 9th IAPR international workshop on document analysis systems (DAS'10), Boston. ACM, New York, pp 97–104. doi:http://doi.acm.org/10.1145/1815330.1815343, URL: http://doi.acm.org/10.1145/1815330.1815343
26. Kanai J, Rice SV, Nartker TA, Nagy G (1995) Automated evaluation of OCR zoning. IEEE Trans Pattern Anal Mach Intell 17:86–90. doi:http://doi.ieeecomputersociety.org/10.1109/34.368146
27. Kanungo T, Haralick RM, Stuezle W, Baird HS, Madigan D (2000) A statistical, nonparametric methodology for document degradation model validation. IEEE Trans Pattern Anal Mach Intell 22:1209–1223. doi:http://dx.doi.org/10.1109/34.888707, URL: http://dx.doi.org/10.1109/34.888707
28. Khosravi H, Kabir E (2007) Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern Recognit Lett 28:1133–1141. doi:10.1016/j.patrec.2006.12.022, URL: http://dl.acm.org/citation.cfm?id=1243503.1243603
29. Kim DW, Kanungo T (2002) Attributed point matching for automatic groundtruth generation. Int J Doc Anal Recognit 5:47–66. doi:10.1007/s10032-002-0083-7, URL: http://dx.doi.org/10.1007/s10032-002-0083-7
30. Lee CH, Kanungo T (2003) The architecture of TRUEVIZ: a groundtruth/metadata editing and visualizing toolkit. Pattern Recognit 36(3):811–825. doi:10.1016/S0031-3203(02)00101-2, URL: http://www.sciencedirect.com/science/article/pii/S0031320302001012
31. Liang J, Phillips IT, Haralick RM (1997) Performance evaluation of document layout analysis algorithms on the UW data set. In: Proceedings of the SPIE document recognition IV, San Jose, pp 149–160
32. Liwicki M, Bunke H (2005) IAM-OnDB – an on-line English sentence database acquired from handwritten text on a whiteboard. In: Proceedings of the eighth international conference on document analysis and recognition (ICDAR'05), Seoul. IEEE Computer Society, Washington, DC, pp 956–961. doi:http://dx.doi.org/10.1109/ICDAR.2005.132, URL: http://dx.doi.org/10.1109/ICDAR.2005.132
33. Liwicki M, van den Heuvel C, Found B, Malik M (2010) Forensic signature verification competition 4NSigComp2010 – detection of simulated and disguised signatures. In: International conference on frontiers in handwriting recognition (ICFHR), Kolkata, 2010, pp 715–720. doi:10.1109/ICFHR.2010.116
34. Liwicki M, Malik M, van den Heuvel C, Chen X, Berger C, Stoel R, Blumenstein M, Found B (2011) Signature verification competition for online and offline skilled forgeries (SigComp2011). In: International conference on document analysis and recognition (ICDAR), Beijing, 2011, pp 1480–1484. doi:10.1109/ICDAR.2011.294
35. Lopresti D (2009) Optical character recognition errors and their effects on natural language processing. Int J Doc Anal Recognit 12:141–151. doi:10.1007/s10032-009-0094-8, URL: http://dx.doi.org/10.1007/s10032-009-0094-8
36. Louloudis G, Stamatopoulos N, Gatos B (2011) ICDAR 2011 writer identification contest. In: International conference on document analysis and recognition (ICDAR), Beijing, 2011, pp 1475–1479. doi:10.1109/ICDAR.2011.293
37. Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R (2003) ICDAR 2003 robust reading competitions. In: Proceedings of the seventh international conference on document analysis and recognition (ICDAR'03), Edinburgh, vol 2. IEEE Computer Society, Washington, DC, pp 682–687. URL: http://dl.acm.org/citation.cfm?id=938980.939531
38. MacLean S, Labahn G, Lank E, Marzouk M, Tausky D (2011) Grammar-based techniques for creating ground-truthed sketch corpora. Int J Doc Anal Recognit 14:65–74. doi:http://dx.doi.org/10.1007/s10032-010-0118-4, URL: http://dx.doi.org/10.1007/s10032-010-0118-4

39. Marti UV, Bunke H (1999) A full English sentence database for off-line handwriting recognition. In: Proceedings of the fifth international conference on document analysis and recognition (ICDAR'99), Bangalore. IEEE Computer Society, Washington, DC, pp 705–708. URL: http://dl.acm.org/citation.cfm?id=839279.840504
40. Mihov S, Schulz K, Ringlstetter C, Dojchinova V, Nakova V, Kalpakchieva K, Gerasimov O, Gotscharek A, Gercke C (2005) A corpus for comparative evaluation of OCR software and postcorrection techniques. In: Proceedings of the eighth international conference on document analysis and recognition, Seoul, 2005, vol 1, pp 162–166. doi:10.1109/ICDAR.2005.6
41. Moll M, Baird H, An C (2008) Truthing for pixel-accurate segmentation. In: The eighth IAPR international workshop on document analysis systems (DAS'08), Japan, 2008, pp 379–385. doi:10.1109/DAS.2008.47
42. Mori M, Suzuki A, Shio A, Ohtsuka S (2000) Generating new samples from handwritten numerals based on point correspondence. In: Proceedings of the 7th international workshop on frontiers in handwriting recognition (IWFHR2000), Amsterdam, pp 281–290
43. Mouchere H, Viard-Gaudin C, Kim DH, Kim JH, Garain U (2011) CROHME2011: competition on recognition of online handwritten mathematical expressions. In: International conference on document analysis and recognition (ICDAR), Beijing, 2011, pp 1497–1500. doi:10.1109/ICDAR.2011.297
44. Ntirogiannis K, Gatos B, Pratikakis I (2008) An objective evaluation methodology for document image binarization techniques. In: The eighth IAPR international workshop on document analysis systems (DAS'08), Nara, 2008, pp 217–224. doi:10.1109/DAS.2008.41
45. Okamoto M, Imai H, Takagi K (2001) Performance evaluation of a robust method for mathematical expression recognition. In: International conference on document analysis and recognition, Seattle, p 0121. doi:http://doi.ieeecomputersociety.org/10.1109/ICDAR.2001.953767
46. Ortega-Garcia J, Fierrez-Aguilar J, Simon D, Gonzalez J, Faundez-Zanuy M, Espinosa V, Satue A, Hernaez I, Igarza JJ, Vivaracho C, Escudero D, Moro QI (2003) MCYT baseline corpus: a bimodal biometric database. IEE Proc Vis Image Signal Process 150(6):395–401. doi:10.1049/ip-vis:20031078
47. Paredes R, Kavallieratou E, Lins RD (2010) ICFHR 2010 contest: quantitative evaluation of binarization algorithms. In: International conference on frontiers in handwriting recognition, Kolkata, pp 733–736. doi:http://doi.ieeecomputersociety.org/10.1109/ICFHR.2010.119
48. Perez D, Tarazon L, Serrano N, Castro F, Terrades O, Juan A (2009) The GERMANA database. In: 10th international conference on document analysis and recognition (ICDAR'09), Barcelona, 2009, pp 301–305. doi:10.1109/ICDAR.2009.10
49. Phillips IT, Chhabra AK (1999) Empirical performance evaluation of graphics recognition systems. IEEE Trans Pattern Anal Mach Intell 21:849–870. doi:http://dx.doi.org/10.1109/34.790427, URL: http://dx.doi.org/10.1109/34.790427
50. Phillips I, Chen S, Haralick R (1993) CD-ROM document database standard. In: Proceedings of the second international conference on document analysis and recognition, Tsukuba, 1993, pp 478–483. doi:10.1109/ICDAR.1993.395691
51. Phillips I, Ha J, Haralick R, Dori D (1993) The implementation methodology for a CD-ROM English document database. In: Proceedings of the second international conference on document analysis and recognition, Tsukuba, 1993, pp 484–487. doi:10.1109/ICDAR.1993.395690
52. Plamondon R, Guerfali W (1998) The generation of handwriting with delta-lognormal synergies. Biol Cybern 132:119–132
53. Pletschacher S, Antonacopoulos A (2010) The page (page analysis and ground-truth elements) format framework. In: 20th international conference on pattern recognition (ICPR), Istanbul, 2010, pp 257–260. doi:10.1109/ICPR.2010.72
54. Pratikakis I, Gatos B, Ntirogiannis K (2010) H-DIBCO 2010 – handwritten document image binarization competition. In: International conference on frontiers in handwriting recognition (ICFHR), Kolkata, 2010, pp 727–732. doi:10.1109/ICFHR.2010.118
55. Pratikakis I, Gatos B, Ntirogiannis K (2011) ICDAR 2011 document image binarization contest (DIBCO 2011). In: International conference on document analysis and recognition (ICDAR), Beijing, 2011, pp 1506–1510. doi:10.1109/ICDAR.2011.299

56. Quiniou S, Mouchere H, Saldarriaga S, Viard-Gaudin C, Morin E, Petitrenaud S, Medjkoune S (2011) HAMEX – a handwritten and audio dataset of mathematical expressions. In: International conference on document analysis and recognition (ICDAR), Beijing, 2011, pp 452–456. doi:10.1109/ICDAR.2011.97
57. Rath TM, Manmatha R (2007) Word spotting for historical documents. Int J Doc Anal Recognit 9(2):139–152. doi:10.1007/s10032-006-0027-8, URL: http://dx.doi.org/10.1007/s10032-006-0027-8
58. Rice SV, Jenkins FR, Nartker TA (1996) The fifth annual test of OCR accuracy. Technical report TR-96-01. AInformation Science Research Institute (University of Nevada, Las Vegas)
59. Rusiñol M, Borrís A, Lladós J (2010) Relational indexing of vectorial primitives for symbol spotting in line-drawing images. Pattern Recognit Lett 31:188–201. doi:http://dx.doi.org/10.1016/j.patrec.2009.10.002, URL: http://dx.doi.org/10.1016/j.patrec.2009.10.002
60. Saund E, Lin J, Sarkar P (2009) PixLabeler: user interface for pixel-level labeling of elements in document images. In: Proceedings of the 2009 10th international conference on document analysis and recognition (ICDAR'09), Barcelona. IEEE Computer Society, Washington, DC, pp 646–650. doi:http://dx.doi.org/10.1109/ICDAR.2009.250, URL: http://dx.doi.org/10.1109/ICDAR.2009.250
61. Schomaker L, Thomassen A, Teulings HL (1989) A computational model of cursive handwriting. In: Plamondon R, Suen CY, Simner ML (eds) Computer recognition and human production of handwriting. World Scientific, Singapore/Teaneck, pp 153–177
62. Serrano N, Castro F, Juan A (2010) The RODRIGO database. In: LREC, Valletta
63. Setlur S, Govindaraju V (1994) Generating manifold samples from a handwritten word. Pattern Recognit Lett 15(9):901–905. doi:10.1016/0167-8655(94)90152-X, URL: http://www.sciencedirect.com/science/article/pii/016786559490152X
64. Shafait F (2007) Document image dewarping contest. In: 2nd international workshop on camera-based document analysis and recognition, Curitiba, pp 181–188
65. 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. ACM, New York, pp 113–120. doi:http://doi.acm.org/10.1145/1815330.1815345, URL: http://doi.acm.org/10.1145/1815330.1815345
66. Smith EHB (2010) An analysis of binarization ground truthing. In: Proceedings of the 9th IAPR international workshop on document analysis systems (DAS'10), Boston. ACM, New York, pp 27–34. doi:http://doi.acm.org/10.1145/1815330.1815334, URL: http://doi.acm.org/10.1145/1815330.1815334
67. Solimanpour F, Sadri J, Suen CY (2006) Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language. In: Lorette G (ed) Tenth international workshop on frontiers in handwriting recognition, Université de Rennes 1, Suvisoft, La Baule. URL: http://hal.inria.fr/inria-00103983/en/
68. Suen C, Nadal C, Legault R, Mai T, Lam L (1992) Computer recognition of unconstrained handwritten numerals. Proc IEEE 80(7):1162–1180. doi:10.1109/5.156477
69. Todoran L, Worring M, Smeulders M (2005) The UvA color document dataset. Int J Doc Anal Recognit 7:228–240. doi:10.1007/s10032-004-0135-2, URL: http://dl.acm.org/citation.cfm?id=1102243.1102245
70. Uchida S, Nomura A, Suzuki M (2005) Quantitative analysis of mathematical documents. Int J Doc Anal Recognit 7:211–218. doi:10.1007/s10032-005-0142-y, URL: http://dl.acm.org/citation.cfm?id=1102243.1102248
71. Varga T, Bunke H (2003) Generation of synthetic training data for an HMM-based handwriting recognition system. In: Proceedings of the seventh international conference on document analysis and recognition (ICDAR'03), Edinburgh, vol 1. IEEE Computer Society, Washington, DC, pp 618–622. URL: http://dl.acm.org/citation.cfm?id=938979.939265
72. Viard-Gaudin C, Lallican PM, Binter P, Knerr S (1999) The IRESTE On/Off (IRONOFF) dual handwriting database. In: Proceedings of the fifth international conference on document

analysis and recognition (ICDAR'99), Bangalore. IEEE Computer Society, Washington, DC, pp 455–458. URL: http://dl.acm.org/citation.cfm?id=839279.840372
73. Wang K, Belongie S (2010) Word spotting in the wild. In: Proceedings of the 11th European conference on computer vision: part I (ECCV'10), Heraklion. Springer, Berlin/Heidelberg, pp 591–604. URL: http://dl.acm.org/citation.cfm?id=1886063.1886108
74. Wang J, Wu C, Xu YQ, Shum HY, Ji L (2002) Learning-based cursive handwriting synthesis. In: Proceedings of the eighth international workshop on frontiers of handwriting recognition, Niagara-on-the-Lake, pp 157–162
75. Wang DH, Liu CL, Yu JL, Zhou XD (2009) CASIA-OLHWDB1: a database of online handwritten Chinese characters. In: Proceedings of the 2009 10th international conference on document analysis and recognition (ICDAR'09), Barcelona. IEEE Computer Society, Washington, DC, pp 1206–1210. doi:http://dx.doi.org/10.1109/ICDAR.2009.163, URL: http://dx.doi.org/10.1109/ICDAR.2009.163
76. Yang L, Huang W, Tan CL (2006) Semi-automatic ground truth generation for chart image recognition. In: Workshop on document analysis systems (DAS), Nelson, pp 324–335
77. Yanikoglu BA, Vincent L (1998) Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation. Pattern Recognit 31(9): 1191–1204. doi:10.1016/S0031-3203(97)00137-4, URL: http://www.sciencedirect.com/science/article/pii/S0031320397001374
78. Zhai J, Wenyin L, Dori D, Li Q (2003) A line drawings degradation model for performance characterization. In: Proceedings of the seventh international conference on document analysis and recognition, Edinburgh, 2003, pp 1020–1024. doi:10.1109/ICDAR.2003.1227813

Further Reading

Datasets are only one of the pillars of any performance evaluation system. The other one is everything related to evaluation, mainly metrics and protocols of evaluation. All these are deeply discussed in ▶Chap. 30 (Tools and Metrics for Document Analysis Systems Evaluation) of this book, and the interested reader can find there a detailed discussion about the different alternatives and approaches used in document analysis.

The topic covered in this chapter is very broad and presents many specific particularities for each subfield of document analysis. Thus, we have only been able to include a very generic description of the datasets. For a further description of the datasets, the reader can go either to the chapter of this book devoted to a specific problem (where some of these datasets are further described) either to the specific references cited throughout the chapter and the appendix. In addition, many of these datasets are being used in the international conferences related to document analysis (mainly, ICDAR, DAS, and ICFHR). The proceedings of these conferences can be another good source of information about the datasets and the current state-of-the-art results for each of them.