

Volker Märgner and Haikal El Abed

Contents

Introduction..... 1012

    Document Processing Evaluation..... 1013

    Evaluation of Systems and Modules..... 1015

Evaluation Metrics..... 1016

    Preprocessing Step: Binarization..... 1018

    Page/Document Layout Analysis..... 1021

    Text Segmentation..... 1024

    Character/Text Recognition..... 1025

Evaluation Tools..... 1028

    Ground-Truthing Tools..... 1028

    Ground Truth in Document Analysis..... 1029

    Architecture of an Evaluation Framework..... 1030

    Contests and Benchmarking..... 1032

Conclusion..... 1034

References..... 1034

    Further Reading..... 1036

Abstract

Any development within a specific research field like document analysis and recognition comes along with the need for data and corresponding measurement devices and test equipment. This chapter introduces the basic issues of evaluation methods for different kind of document analysis systems and modules with a special emphasis on tools and metrics available and used today.

This chapter is organized as follows: After a general introduction including general definitions of terms used in document analysis system evaluation and general overviews of evaluation processes in section “Introduction,” different

V. Märgner (✉) • H.E. Abed  
Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig,  
Germany  
e-mail: [maergner@ifn.ing.tu-bs.de](mailto:maergner@ifn.ing.tu-bs.de); [el-abed@ifn.ing.tu-bs.de](mailto:el-abed@ifn.ing.tu-bs.de)

evaluation metrics are discussed in section “[Evaluation Metrics](#).” These metrics cover the different aspects of the Document Analysis Handbook as presented in ► [Chaps. 2](#) (Document Creation, Image Acquisition and Document Quality)–► [8](#) (Text Segmentation for Document Recognition), from image-processing evaluation metrics to special metrics for selected applications e.g., character/text recognition. In section “[Evaluation Tools](#),” an overview of ground-truth file structure and a selection of available ground-truth tools are presented. Performance evaluation tools and competitions organized within the last years are also listed in section “[Evaluation Tools](#).”

---

**Keywords**

Benchmark • Competition • Evaluation • Evaluation tools • Ground truth • Metrics • Quality measure

---

## Introduction

Recognition tasks usually transform signals – one-, two- or more dimensional – into symbols of a known alphabet or vocabulary/lexicon. From the very beginning the most important measurement of the quality of such systems is based on a comparison of the ground truth and the system output symbol string. At first for such tasks, metrics and tools have to be developed. Very soon it became clear that obviously the comparison with the output only of such complex systems represents a general approach and does not take into account different aspects which may play an important role in the transformation process from an image to a symbol string. Especially systems for dedicated applications or single modules development of document analysis systems necessitate a more decided quality measurement. As a consequence different metrics and tools were developed on color, grey, or binary value of the image and on the object shape.

Before a short introduction to system evaluation starts, the definitions of the terms used in system development and evaluation are given:

**Assessment:** “**Assessment** is the collection of relevant information that may be relied on for making decisions. **Evaluation** is the application of a standard and a decision-making system to assessment data to produce judgments about the amount and adequacy of the learning that has taken place.” Too often these processes are confused ...it is said assess, but meant evaluate ...or the term evaluation is used, when really assessment [11] is done.

**Evaluation:** As defined by the American Evaluation Association (<http://www.eval.org/>), evaluation involves assessing the strengths and weaknesses of programs, policies, personnel, products, and organizations to improve their effectiveness.

**Benchmarking:** For the term benchmarking the Oxford Dictionaries gives the following definition: Benchmarking means a standard or point of reference against which things may be compared. This is for example a problem designed to evaluate the performance of a computer system.

**Ground truth:** Ground truth is a term first used in cartography, meteorology, analysis of aerial photographs, satellite imagery, and a range of other remote sensing techniques in which data are gathered at a distance. Ground truth refers to information that is collected “on location.” In document analysis and recognition techniques, the term “ground truth” refers to the correct meaning or classification of a data set, used for the training and testing for supervised learning techniques.

**Comparison:** Means the act of comparing. It is an examination of two or more objects with the view of discovering the resemblances or differences. (<http://www.definitions.net/definition/comparison>)

**Contest/Competition:** These terms mean the act of seeking, or endeavoring to gain, what another is endeavoring to gain at the same time. (<http://www.definitions.net/definition/competition>)

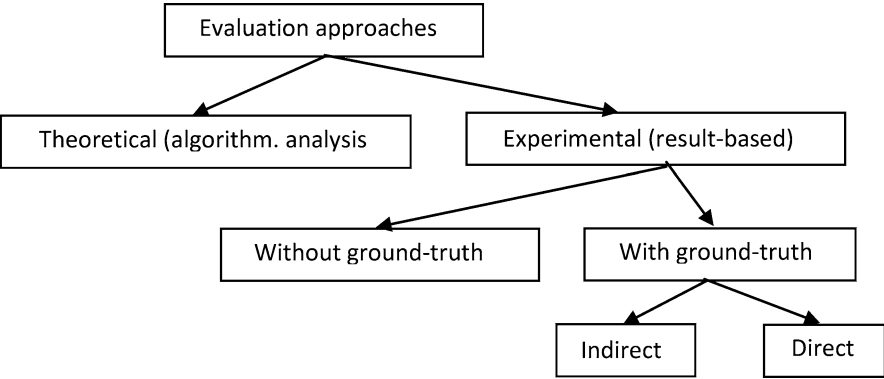
## Document Processing Evaluation

Any document analysis and recognition system need quality evaluation or benchmarking as the increasing quality and diversity of document analysis systems have made it more and more difficult to compare systems. Moreover, as the complexity of the systems has increased with the effect that modifications within one module – even if just one parameter is modified – may often lead to an unpredictable behavior towards other modules. The increasing amount of ready-to-use algorithms and the exploitation of new application fields for document analysis are making it rather difficult to find a suitable configuration. In conclusion, both, a detailed qualitative and quantitative failure analysis are needed for further improvement.

When making an evaluation, it can be distinguished between two different objectives:

- Benchmarking for the system user: In this case, only the final results (e.g., the ASCII text output of an OCR system (Optical Character Recognition) or a set of categories output of a document categorization system) are of interest. There is no motivation to look in more detail at internal modules.
- Benchmarking for the system developer: In this case a detailed failure analysis is necessary. This requires focusing not only on final system output but also on intermediate results, i.e., on the output of single system modules. That means a module’s output data must be accessible for an evaluation, e.g., comparison with corresponding ground-truth data.

Evaluation approaches itself may be classified with the scheme shown in Fig. 30.1. It can be distinguished between theoretical and experimental approaches. Using the theoretical approaches, the algorithm implemented will be analyzed in order to derive its behavior. The input data are assumed to be composed of ideal data degraded by noise. Input data degradation is propagated by analytical description throughout the algorithm. Approaches like these theoretical ones are mainly used in the domain of low-level computer vision algorithms. Such approaches would hardly be applicable for segmentation modules as part of a document analysis system as



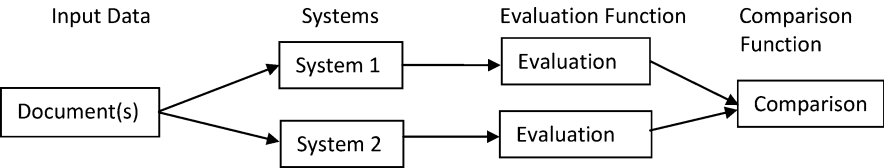
**Fig. 30.1** Classification of evaluation approaches

the implemented algorithms are quite more complex to that of image preprocessing modules.

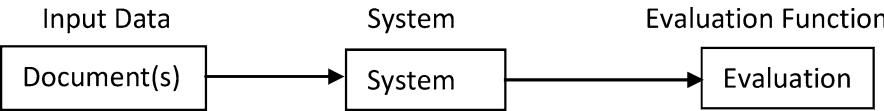
On the contrary, experimental evaluation without ground truth considers a system or a module characterized as a black box. The evaluation is based on the system or module output only. Approaches like this are much more pragmatic and especially they are independent of the underlying algorithm. Using evaluation methods without ground truth, it is possible to implement online measures which can be used to evaluate a specific aspect of quality. In this context online means that the system itself can calculate the measure of assessing the quality of its results, ground truth is not available. But when making a detailed and comprehensive quality evaluation, this methodology cannot be sufficient.

When using ground truth, quality evaluation means to make a comparison of the output with the ground truth. It depends on the type of data whether the comparison is simple or more complicated. It can easily be seen that the comparison of segmentation results is more complicated than, e.g., the comparison of character classification results because segmentation results are neither isolated like character patterns nor of a symbolical kind.

Not least because of this, there exists the idea that comparison may not be performed on the basis of the direct results but on the basis of indirect results. That means a further processing of the results has to be done. For example, the textual results which a document analysis system generates may be compared with the plain ground-truth text even to evaluate the zoning quality of a system. The comparison may be done by means of a string-matching algorithm. A great advantage of this approach is that ground truth is only needed on a textual level and not on the segment level. But the disadvantage is that an exact quality evaluation cannot be performed, due to the loss of information when the segments are converted to a stream of character symbols. This disadvantage does not exist when an evaluation is made on the basis of direct results, say, of the segment data itself.



**Fig. 30.2** Comparison of two systems



**Fig. 30.3** Evaluation of one system

Besides the question how to perform the evaluation, there is the task to provide test data. There are two extremes: the collection of real data and, since this is often a very laborious task, the generation of synthetic data.

**Evaluation of Systems and Modules**

Evaluation basically pursues three aims:

- Improvement and quality control of a system
- Comparison of two systems
- Definition of the absolute performance of a system

As the first aim tends to increase or maintain a system’s performance, the second compares systems to get the information which one is the better with respect to a certain task. The third aim is the definition of the absolute performance of a system. Figures 30.2 and 30.3 show block diagrams of different types of document analysis systems evaluation.

Now let us have a closer look to the blocks of the evaluation system: input data, system, evaluation function, and comparison function. The input data, in our case scanned documents, need to be representative for the population of the system’s task. The system itself is a description of the vaguely defined task but not necessarily of a completely defined algorithm. This is important to have in mind when an ideal system and not a real system is in consideration. That means a system may be described through the desired output only. This output is on the same symbolic level as the ground truth. The evaluation function assigns a value to the output, e.g., based on a comparison of the output symbol sequence with the corresponding ground truth. The evaluation function is less decisive for OCR or computer systems, for instance, but a crucial factor in areas like image processing or segmentation. Finally the comparison function compares the evaluation values of different systems. Evaluation methods may be defined by a human expert, the subsequent system, or a comparison with ground truth. The evaluation and the

comparison functions have to use metrics useful to measure most important system core module features. In the following evaluation metrics, at first general and later specific document-processing tasks are presented.

## Evaluation Metrics

Image quality in general plays an important role in document analysis and recognition, as all tasks applied to the document image need a certain image quality to be successful. But the definition of image quality metrics is difficult as it usually needs a ground-truth image as a basis to calculate a quality measure. In addition to the often used peak signal-to-noise ratio (PSNR) metric as a simple statistical measure using pixel differences, especially for image coding human visual perception is integrated into the metric, for example, the knowledge that in an image noise close to a high contrast edge has a different influence to the image quality than on a part of the image with constant color. For document analysis systems the quality metric has to be aligned to the typical content and also to the typical noise in scanned documents. Performing a system evaluation needs a measure for quantifying the performance of the whole system or module or even parts of them. A metric is used as such an objective measure.

Image quality metrics are widely used for image coding quality measure. In case of image coding both the original and the coded image are available, the quality loss introduced by the coding can be measured by comparing both images. The image preprocessing task is more difficult as the original image which is the goal of any preprocessing task is usually not known.

A metric to measure the distance between original and distorted image is the peak signal-to-noise ratio (PSNR). PSNR is the ratio between the maximum possible power of a signal, in our case the signal is an image, and the power of the noise affecting the quality of the signal (image). For image processing this metric is primarily used to measure the quality of a lossy image compression codec, where the compression introduces some noise into the original image. PSNR is defined by using the mean square error (MSE). Given an image of size  $N_x \times N_y$ , the noise-free original image is given by  $I_1(x, y)$  and the noisy version by  $I_2(x, y)$ .

The MSE is defined as

$$MSE = \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} (I_1(x, y) - I_2(x, y))^2}{N_x \times N_y} \quad (30.1)$$

And the corresponding PSNR as

$$PSNR = 10 \cdot \log_{10} \left( \frac{\text{Max}_I^2}{MSE} \right) \quad (30.2)$$

$\text{Max}_I$  is the maximum possible pixel value appearing in the image. In case of an 8 bit per pixel grey value image, this value is 255. For color images with RGB presentation of the colors, the definition is the same except the MSE is the sum of all 3 (R, G, and B) channels squared differences divided by image size and by 3. To use this measure for image preprocessing, e.g., for noise reduction, a noise free version of the image is needed as reference. The measure of PSNR uses this image and the preprocessed one. The best possible image preprocessing finally results in an image equal to the noise-free version of the image with  $\text{MSE} = 0$  and an infinite PSNR. Typical preprocessing steps result in an  $\text{MSE} > 0$  and  $\text{PSNR} < \text{infinity}$ .

Metrics used in document analysis systems evaluation often originate from the field of information retrieval or information extraction. This is understandable as the main goal of document analysis can be seen as an information extraction process also, since the information content out of a document has to be extracted.

Most important and widely used metrics in the field of information extraction are recall (R) and precision (P). Information extraction – the task for the extraction of a certain word out of a text-deals with completeness of retrieval (all objects are extracted) and the purity of retrieval (only correct objects are retrieved). Recall is defined as correctly retrieved objects out of all objects in the data set, which gives a measure for the completeness of retrieval. Precision is defined as correctly retrieved objects out of all objects retrieved, which gives a measure for the purity of retrieved data.

Recall and precision consider two different aspects of retrieval systems. One may be of more interest in the precision if wrongly retrieved objects are highly undesirable; others have more problems if objects present in the data set are missed. To evaluate the overall quality of a system, a further metric is defined taking into account both recall and precision. Basic idea is to bring the two measures together and calculate the weighted mean of both. The weight is used to take different importance of recall and precision into account. As precision and recall are ratios with the same denominator but different enumerator, the best mean is not the often used arithmetic mean but the harmonic mean, defined as

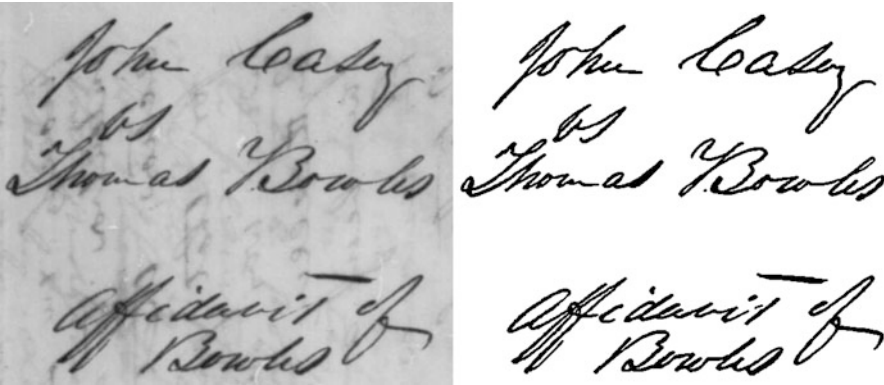
$$H = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R} = FM \quad (30.3)$$

This unweighted harmonic mean is called F-measure (FM) and used as a metric for different applications in data retrieval.

In the following metrics to evaluate document preprocessing, page layout analysis, text segmentation, and recognition are presented. It can be seen that the aforementioned metrics are used in many different applications in the document processing field.



**Fig. 30.4** Image of a part of a book page, *left side* color image, *right side* corresponding binary ground-truth (GT) image



**Fig. 30.5** Image of a part of a letter, *left side* grey scale, *right side* corresponding binary ground-truth (GT) image

**Preprocessing Step: Binarization**

Noise reduction is a commonly used preprocessing technique with usually not a strong influence to the subsequent processing steps. Binarization is a preprocessing step in many cases essential for the next steps in the processing chain. Using a proper binarization is therefore an extremely important decision for the system performance. The comparison between binarization methods is based on several metrics described in the following. These metrics were used in document image binarization contests and first defined in [31]. The evaluation of binarization methods is basically based on comparison between the binary image and the corresponding ground-truth image (GT). In both images black pixels (equal to 1) are classified as foreground and white pixels (equal to 0) are classified as background (►Chap. 4 (Imaging Techniques in Document Analysis Processes)).

Figures 30.4 and 30.5 show a small part of a printed and a handwritten document together with the corresponding ground-truth binary image. The images are examples from the H-DIBCO 2009 Handwritten Document Image Binarization Contest



held during the International Conference on Document Analysis and Recognition in Barcelona 2009 [12].

### F-Measure (FM)

A very important metric often used for binarization evaluation is the F-measure (*FM*) (see Eq. 30.3 for definition) which is calculated pixel-wise in case of binarization. The metric *FM* is calculated according to *Recall* (Eq. 30.5) and *Precision* (Eq. 30.6) as it is shown in Eq. 30.4.

$$FM = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (30.4)$$

$$Recall = \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} GT(x, y) \times B(x, y)}{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} GT(x, y)} \times 100 \quad (30.5)$$

$$Precision = \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} GT(x, y) \times B(x, y)}{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} B(x, y)} \times 100 \quad (30.6)$$

In both Eqs. 30.5 and 30.6, *GT* and *B* denote the ground-truth and the binary image respectively with foreground (black) pixels coded with 1 and background (white) pixels coded with 0.

It is easy to see that in this case FM is the harmonic mean of the ratio of number of same pixels in the ground-truth binary image and the binary image under test to pixels in the ground-truth binary image in case of recall and the binary image under test in case of precision.

### Peak Signal-to-Noise Ratio (PSNR)

As second metric for the evaluation of binarization methods the aforementioned peak signal-to-noise ratio (PSNR) (see Eq. 30.2) is used. *PSNR* measures the difference of the binary image under test compared to the ground-truth binary image and is calculated as follows:

$$PSNR = 10 \cdot \log_{10} \left( \frac{C^2}{MSE} \right) \quad (30.7)$$

*C* is a value showing the maximal difference between foreground and background pixel intensities; this value is set to 1 in the case of binary images. The mean square error *MSE* is calculated by Eq. 30.8.

$$MSE = \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} (GT(x, y) \times B(x, y))^2}{N_x \times N_y} \quad (30.8)$$

$GT$  and  $B$  denote the ground-truth and the binary images respectively with foreground (black) pixel coded with 1 and background (white) pixel coded with 0.

This objective measure takes into account different pixel values without consideration of the subjective visual perception of the difference.

### Negative Rate Metric (NRM)

A more specific metric is the Negative Rate Metric (NMR).  $NRM$  represents the relationship between the ground-truth ( $GT$ ) and the binary image ( $B$ ) pixels.  $\#TP$ ,  $\#FP$ ,  $\#FN$ , and  $\#TN$  denote the number of true positive ( $GT = 1$  and  $B = 1$ ), false positive ( $GT = 0$  and  $B = 1$ ), false negative ( $GT = 1$  and  $B = 0$ ), and true negative ( $GT = 0$  and  $B = 0$ ), respectively.  $NRM$  is calculated using Eq. 30.9.

$$NRM = \frac{\frac{\#FN}{\#FN + \#TP} + \frac{\#FP}{\#FP + \#TN}}{2} \quad (30.9)$$

$NRM$  is a value calculated as arithmetic mean of foreground pixels changed to background in relation to all foreground pixels and background pixel changed to foreground in relation to all background pixels.

### Distance Reciprocal Distortion Metric (DRD)

The Distance Reciprocal Distortion Metric (DRD) takes into account not only the pixel value itself but also the value of pixels close to the one changed its color.  $DRD$  has been used also to measure the subjective visual distortion in binary document images [22]. It properly correlates with the human visual perception, and it measures the distortion for all  $S$  pixels which color has changed from white to black or vice versa (flipped pixels) as follows:

$$DRD = \frac{\sum_{k=1}^S DRD_k}{N_{UBN}} \quad (30.10)$$

where  $DRD_k$  is the distortion of the  $k$ th flipped pixel and is calculated using a  $5 \times 5$  normalized weight matrix  $W_{N_m}$  (Table 30.1) as defined in [22], where each neighboring pixel is weighted by a reciprocal of the distance to the flipped pixel in the center.  $DRD_k$  equals to the weighted sum of the pixels in the  $5 \times 5$  block of the ground-truth  $GT$  that differ from the centered  $k$ th flipped pixel at  $(x, y)$  in the binarization result image  $B$  (Eq. 30.11).

**Table 30.1** Normalized weight matrix  $W_{Nm}$ 

0.0256	0.0324	0.0362	0.0324	0.0256
0.0324	0.0512	0.0724	0.0512	0.0324
0.0362	0.0724	0	0.0724	0.0362
0.0324	0.0512	0.0724	0.0512	0.0324
0.0256	0.0324	0.0362	0.0324	0.0256

$$DRD_k = \sum_{i=-2}^2 \sum_{j=2}^2 |GT_k(i, j) - B_k(x, y)| \times W_{Nm}(i, j) \quad (30.11)$$

Finally,  $NUBN$ , the denominator in Eq. 30.10, is the number of the nonuniform (not all black or all white pixels)  $8 \times 8$  blocks in the GT image.

$DRD$  is a value where each pixel with different value to the ground truth is not only counted but weighted by the number of pixels in the neighborhood with different value than the centered ground-truth pixel value. The normalization factor is the sum of all nonuniform  $8 \times 8$  pixels in the ground-truth image.

### Misclassification Penalty Metric (MPM)

Finally the Misclassification Penalty Metric (MPM), a metric which uses the mean position change of contour points between ground-truth and binary image. As defined in Eq. 30.12,  $MPM$  measures a mean distance between contour points of the ground-truth (GT) and the binary image ( $B$ ).  $d_{FN}^a$  is here defined as the distance between the  $a$ th false negative ( $GT = 1$  and  $B = 0$ ) and the closest contour pixel of the binary image.  $d_{FP}^b$  is defined as the distance between the  $b$ th false positive ( $GT = 0$  and  $B = 1$ ) and the closest contour pixel of the ground-truth image. The normalization factor  $D$  is the overall sum of the pixel-to-contour distances of the ground-truth object.

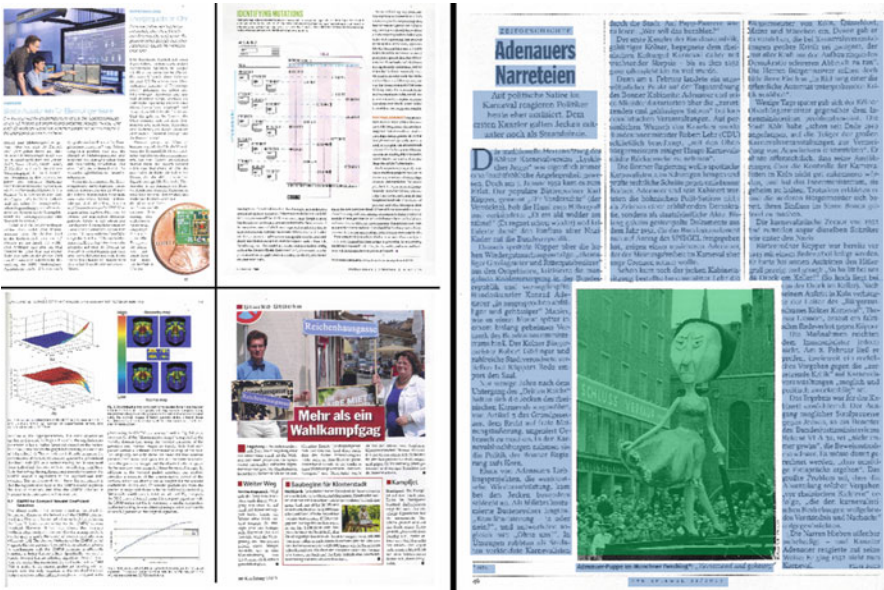
$$MPM = \frac{\sum_{a=1}^{\#FN} d_{FN}^a + \sum_{b=1}^{\#FP} d_{FP}^b}{D} \quad (30.12)$$

The  $MPM$  metric takes into account that the mean distance of contour points in GT and B images is a good quality measure as a small distance is a less important error than a large distance.

All these different metrics used for the evaluation of binarization results show us a large variety. It depends on the application which metric is the best to use as different errors result in different penalties. In the following some of the most important metrics used for page layout evaluation are presented.

## Page/Document Layout Analysis

Document layout analysis aims to the segmentation of a document page into logical areas (see ► [Chaps. 5](#) (Page Segmentation Techniques in Document Analysis) and



**Fig. 30.6** Sample pages for layout analysis on the *left* and an example for a ground-truth layout analysis of a magazine page on the *right*. (These types of pages are used in [3] for layout analysis performance evaluation.)

►6 (Analysis of the Logical Layout of Documents)). The image of a scanned page is used first to analyze the physical structure of the page image. The physical structure is subsequently transformed into the page logical structure which is essential to read (understand) a document’s content. Metrics, used to evaluate the performance of page layout analysis, have to deal with the comparison of image areas and not so much with pixel comparison and the corresponding values as it is the case for evaluation of image quality, preprocessing, and binarization as seen in section “Preprocessing Step: Binarization.”

Figure 30.6 shows some examples of pages with different layouts. As can be seen the layout differs a lot. The layout depends a lot on the document type (e.g., scientific journal or boulevard magazine). The ground truth is given by assigning each physical block to a logical part of the document, e.g., text, image, and graphic, as shown in the right part of Fig. 30.6. In the following metrics to measure the performance of page layout, analysis systems are presented.

The performance evaluation method presented here is that one used in ICDAR 2005 Page Segmentation Competition [2] which is based on counting the number of matches between the entities detected by the algorithm and the entities in the ground-truth image [8, 29, 30]. From these results a global *MatchScore* table for all entities is build whose values are calculated according to the intersection of the ON pixel sets of the result and the ground truth (a similar technique is used in [40]). Let  $I$  be the set of all image points,  $G_j$  the set of all points inside the ground-truth region  $j$ ,  $R_i$  the set of all points inside the result region  $i$ , and  $T(s)$  a function that counts the elements of set  $s$ . Table *MatchScore*( $i, j$ ) represents the matching results

of the ground-truth region  $j$  and the result region  $i$ . The global *MatchScore* table for all entities is defined as:

$$MatchScore(i, j) = \frac{T(G_j \cap R_i \cap I)}{T((G_j \cup R_i) \cap I)} \quad (30.13)$$

If  $N_i$  is the count of ground-truth elements belonging to entity  $i$ ,  $M_i$  is the count of result elements belonging to entity  $i$ , and  $w_1, w_2, w_3, w_4, w_5, w_6$  are predetermined weights, the detection rate and recognition accuracy for entity  $i$  can be calculated as follows:

$$DetectRate_i = w_1 \frac{one2one_i}{N_i} + w_2 \frac{g\_one2many_i}{N_i} + w_3 \frac{g\_many2one_i}{N_i} \quad (30.14)$$

$$RecognAccuracy_i = w_4 \frac{one2one_i}{M_i} + w_5 \frac{d\_one2many_i}{M_i} + w_6 \frac{d\_many2one_i}{M_i} \quad (30.15)$$

where the entities  $one2one_i, g\_one2many_i, g\_many2one_i, d\_one2many_i$ , and  $d\_many2one_i$  are calculated from *MatchScore* table (defined in Eq. 30.13) following the steps for every entity  $i$  described in [5]. Here  $one2one_i$  is the number of correctly detected elements,  $g\_one2many_i$  is the number of cases where more than one detected element match with one ground-truth entity,  $g\_many2one_i$  is the number of cases where one detected element match with more than one ground-truth entity,  $d\_one2many_i$  and  $d\_many2one_i$  are equivalent cases with detected and ground-truth entity interchanged.

Based on this definition of detection rate and recognition accuracy, the Entity Detection Metric (EDM) is defined which represents a performance metric for detecting each entity. *EDM* is based on the harmonic mean of detection rate and recognition accuracy as defined in Eqs. 30.14 and 30.15. The Entity Detection Metric ( $EDM_i$ ) of entity  $i$  is defined as follows:

$$EDM_i = \frac{2 \times DetectRate_i \times RecognAccuracy_i}{DetectRate_i + RecognAccuracy_i} \quad (30.16)$$

Finally a global performance metric using all entities detected is defined by the combination of all values of detection rate and recognition accuracy. If  $I$  is the total number of entities and  $N_i$  is the count of ground-truth elements belonging to entity  $i$ , then by using the weighted average (arithmetic mean) for all  $EDM_i$  values, the Segmentation Metric (SM) is defined with the following formula:

$$SM = \frac{\sum_I N_i EDM_i}{\sum_I N_i} \quad (30.17)$$

## Text Segmentation

Once the page layout is finished, all recognized text areas have to be prepared for the recognition process (see ►Chap. 8 (Text Segmentation for Document Recognition)). In this step text blocks have to be segmented into lines. Evaluation of the text segmentation results is again based on comparison of ground-truth areas with segmented areas. Several evaluation metrics are used all based on ground-truth images where the output label image is compared to its corresponding label ground truth.  $S_i$  denotes the  $i$ th text-line detected with the segmentation method under evaluation, where  $i \in \{1; \dots; N_r\}$  and  $N_r$  is the number of the extracted lines.  $G_j$  denotes the  $j$ th text-line in the ground truth image, where  $j \in \{1; \dots; N_g\}$  and  $N_g$  is the number of text-lines in the input image.

### F-Measure (FM)

For line segmentation again the F-measure (FM) as it was used for the binarization step (compare section “F-measure (FM)”) is again a very important metric. FM here depends on the match  $\text{MatchScore}(i,j)$  between  $S_i$  and  $G_j$ . This score is slightly modified to that one presented in section “Page/Document Layout Analysis.” A match is counted if the  $\text{MatchScore}(i,j)$  is equal or higher than a threshold  $Ta$ .  $Ta$  is fixed based on experiments.  $No$  is called number of counted match (Eq. 30.18).

$$No = \# \text{MatchScore}(i,j) \geq Ta \quad (30.18)$$

FM is calculated using precision here called detection rate  $DR$  ( $DR = No/N_g$ ) and recall here called recognition accuracy  $RA$  ( $RA = No/N_r$ ) as follows:

$$FM = \frac{2 \times DR \times RA}{DR + RA} \quad (30.19)$$

### Error Rate (U)

The error rate ( $U$ ) is a metric, used to integrate all different types of misclassification into one value  $U$ . The error rate  $U$  is calculated using the following equation:

$$U = \frac{\#(\psi_1 Ns + \psi_2 Nm + \psi_3 Ne + \psi_4 Np)}{\#Ng} \quad (30.20)$$

where  $Ns$  represents the number of split regions. A region  $G_j$  is a split region if it coincides at minimum with two different  $S_i$  in such a way that the percentage of the  $G_j$  pixels in each segmentation region  $S_i$  is equal or higher than a predefined threshold  $Tb$ .

$Nm$  is a value for the number of merged regions. A region  $S_q$  is called a merged region if it coincides at minimum with two different  $G_j$ , such as the percentage of each  $G_j$  region in  $S_i$  is equal or higher than a predefined threshold  $Tb$ .

$Ne$  is the number of missed regions, where  $G_j$  is a missed region if it does not coincide with any  $S_i$ . Finally  $Np$  is the number of partially missed regions, where  $G_j$

is called a partially missed region, if the percentage of detected pixels of  $G_j$  in  $S_i$  is less than a predefined threshold  $Tb$ .

The weight values  $\psi_1$ ,  $\psi_2$ ,  $\psi_3$ , and  $\psi_4$  have to be fixed according to test on training pages. The same has to be done for the threshold  $Tb$ .

For the segmentation of lines into words and words into characters, similar metrics are used.

## Character/Text Recognition

Character/text recognition is a main module of a document recognition system as with this module, the image of a character, a word, or a text is transferred into the symbolic representation of a character, word, or text. The performance of this part is essential for the performance of the whole system. Comparing different character recognition systems with each other is not really simple as such systems need all modules mentioned in the paragraphs before, e.g., preprocessing, layout analysis, and segmentation. The evaluation of printed character recognition therefore has to take into account the presence of all modules. Let us divide the problem into the following different steps:

1. The input of the recognition system is the image of a single character.
2. The input of the system is the image of a single word.
3. The input of the system is the image of a single line.
4. The input of the system is the image of a page.

Except for case 1 for all other cases, an error may come from recognition or segmentation error. In the following lists showing different types of errors which may occur as output of a character/text recognition system for each of the above mentioned, four cases of input are given:

### Case 1 errors:

**Substitution** – recognizing one character as another

### Case 2 errors:

**Substitution** – recognizing one character as another. This often happens for structurally close characters.

**Deletion** – ignoring a character because it is regarded as noise or as part of another character.

**Insertion** – one symbol is recognized as two symbols or a noise is recognized as character.

**Rejection** – either the system can't recognize a character or it is not sure in its recognition.

**Case 3 errors:** In addition to errors on character level from case 2, errors on word level may occur:

**Word substitution** – recognizing one word as another.

**Word deletion** – ignoring a word because it is regarded as noise or as part of another word.

**Word insertion** – one word is recognized as two words or noise is recognized as a word.

**Word rejection** – either the system can't recognize a word or it is not sure in its recognition.

**Case 4 errors:** In addition to errors on character level (case 2) or word level (case 3), errors on page level may occur:

**Line deletion** – ignoring a line because it is regarded as noise or as part of another line

**Line insertion** – one line is recognized as two lines or noise is recognized as a line

It is easy to understand that compared to case 1 errors, all other errors depend on classification and segmentation, as only in case 1 the system “knows” that the input is a character. In the following most common metrics used for a character recognition task are presented.

### Character Error Rate (CER)/Character Recognition Rate (CRR)

The error rate on character level CER is the percentage of number of wrong recognized characters #CERR in relation to the total number of characters #CHR in the ground-truth text. Wrong recognized characters mean all types of errors: substitution, deletion, insertion, and rejection on character level.

$$CER = \frac{\#CERR}{\#CHR} \quad (30.21)$$

The recognition rate on character level CRR has a similar definition, but here the number of correct recognized characters #CRC is set into relation to all characters.

$$CRR = \frac{\#CRC}{\#CHR} \quad (30.22)$$

### Word Error Rate (WER)/Word Recognition Rate (WRR)

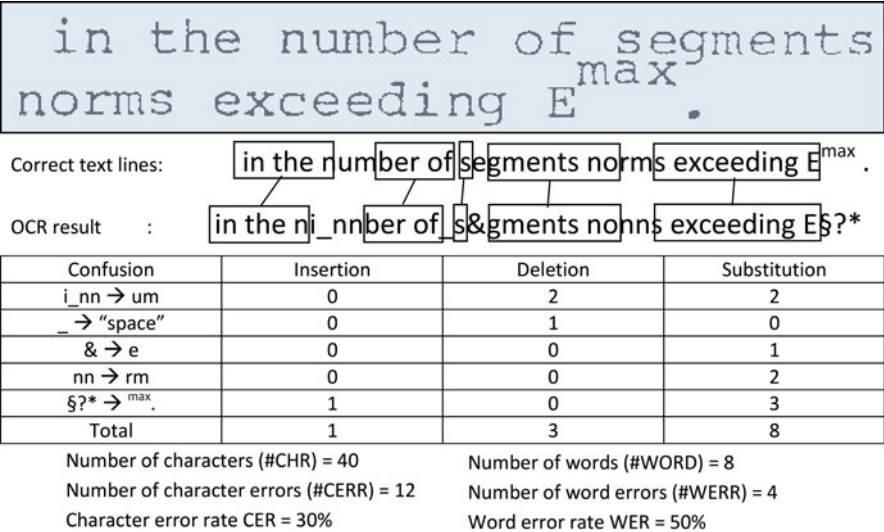
The word error rate and word recognition rate are in the same way defined as for characters. The number of wrong recognized words is #WERR, the number of correct recognized words is #CRW, and the number of words in the ground-truth text #WORD. With these definitions WER and WRR are given by

$$WER = \frac{\#WERR}{\#WORD} \quad (30.23)$$

$$WRR = \frac{\#CRW}{\#WORD} \quad (30.24)$$

Figure 30.7 gives an example of a page segment with recognized text and different types of errors.





**Fig. 30.7** From *top to bottom*: Part of a text page image, ground-truth text on this image, recognition (OCR) result, errors classified and three types of errors: Insertion, deletion, and substitution

**Accuracy (or Recall)**

Accuracy (or recall, see also beginning of section “[Evaluation Metrics](#)” ►[Chap. 2](#) (Document Creation, Image Acquisition and Document Quality)) is defined as the ratio of the number of correctly recognized characters (#CCR) to the number of characters in the ground-truth data (#CHR)

$$\text{Accuracy (or recall)} = \frac{\#CCR}{\#CHR} \tag{30.25}$$

Accuracy (or recall) is in the same way defined for any entity recognized, e.g., number of correctly recognized text frames divided by total number of text frames in data set under research.

**Precision**

Precision is defined as the ratio of correctly recognized number of characters #CCR to number of characters from OCR output (#COCR)

$$\text{Precision} = \frac{\#CCR}{\#COCR} \tag{30.26}$$

Precision is in the same way defined for any entity recognized, e.g., number of correct recognized words divided by total number of words returned by the recognizer for the whole data set under research.

### Cost

The cost value is defined as the weighted sum of editing operations needed to correct the recognized character string. Operations are deletion (**D**), insertion (**I**), and substitution (**S**). The number of each operation is weighted by a predefined factor. The cost is defined as follows:

$$\text{Cost} = w_D \times \#D + w_I \times \#I + w_S \times \#S \quad (30.27)$$

where  $w_D$ ,  $w_I$ ,  $w_S$  are weights which are set to a predefined value given dependent to the text content.

---

## Evaluation Tools

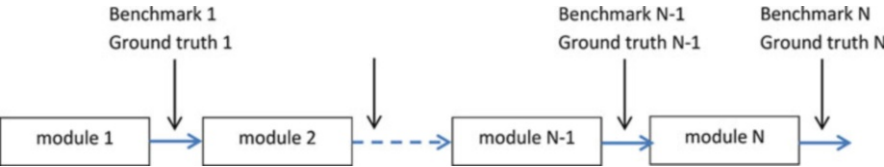
Performance evaluation is one of most important steps in a document-processing system. The complexity of this step depends on the system modules and outputs. Last 20 years, different performance evaluation tools were developed and used as “standard” evaluation tools for specific applications. In this section, in the first part ground-truthing principals and tools are presented and in the second part an overview about competitions and benchmarking campaigns.

### Ground-Truthing Tools

In this section, ground truth is defined in terms that were introduced before. In document analysis, GT denotes the optimal output of a system. This term will be adopted and defined for benchmarking in general. To define GT, one should consider that optimal always asks for a reference: “optimal with respect to what?”

For example, in document analysis, GT was mainly defined for OCR output. For the OCR output, GT was simply the sequence of characters that the human would recognize from the input image. So in this case, a person determined what is optimal. Suppose, for a handwritten character GT has to be found. One person recognizes an “A” and another one an “H.” In this case, the question is “optimal with respect to whom?”

When defining GT at the output of some interior module of a system, the optimal output of the module will depend on the subsequent modules following to the final output of the overall system. Thus GT at any interface within the system has to be defined with respect to the following subsystem (Fig. 30.8).



**Fig. 30.8** Ground truth can be defined anywhere in a system

### Ground Truth in Document Analysis

In the previous sections, GT in general was discussed and its role in document analysis explained. In general, it is used for comparison with the actual output of a system to benchmark. In every goal of benchmarking, it has to be known. Thus, it is essential for benchmarking. It was already stressed how GT is gained for OCR systems, but it is still unclear how it can be found for image-processing systems in document analysis without any classification. Thus, in the following preprocessing steps will be in the focus.

Basically, there are several possibilities to determine GT:

- Definition by a person using adequate rules
- Synthetic generation
- Iterative approximation based on an appropriate assessment function

In the following a short overview of the usage of GT as mentioned in literature and our approach will be given. Haralick, Jaishima, and Dori [14] claim that synthetic imagery is necessary for performance measurement because of the limited real images with GT. Their method requires a system existing in an algorithmic form for which only a set of parameters is still undetermined. Moreover, an ideal world of input data has to be defined which will be propagated through the system to constitute GT. Afterwards the input images are perturbed according to an input perturbation model and the output is compared to GT.

Another possibility is asking a human to determine GT, i.e., to replace the assessment function. Lee, Lam, and Suen [20] defined GT by asking five experts and five nonexperts to draw an optimal skeleton. From these images, some kind of average skeleton was computed.

As third method to define GT, an iterative approximation of GT using an assessment function is presented. The idea is to give an input to the assessment function and to judge its closeness to GT by looking at the output. The former output is taken as a reference for the following modifications of the input which is treated as described above. GT is found if the assessment function equals zero. The assessment could be, for example, a human or the subsequent system as mentioned above.

**Table 30.2** Overview of published ground-truthing tools

Tool	Description	Topic(s)
INSEGD [7], 1995	Tools for document database generation	Document
Pink Panther [40], 1998	Ground-truthing and benchmarking document page segmentation environment	Page segmentation
TrueViz [19], 2001	TrueViz (Ground TRUth/metadata Editing & VISualiZing Toolkit) is a tool for visualizing and editing ground truth and metadata for OCR	OCR, Document
PerfectDoc [39], 2005	Ground-Truthing Environment for Complex Documents	Document
DIAU system [15], 2007	Automatic Ground-truth Generation for Document Image Analysis and Understanding	Document analysis
PixLabeler [35], 2009	Tool for labeling elements in a document image at a pixel level	Document
GEDI [10], 2010	Ground-truthing Environment for Document Images	Document
CBRDIA [6], 2011	Administrative Document Analysis and Structure	Document
Aletheia [9], 2011	Document Layout and Text Ground-truthing System for Production Environments	Document

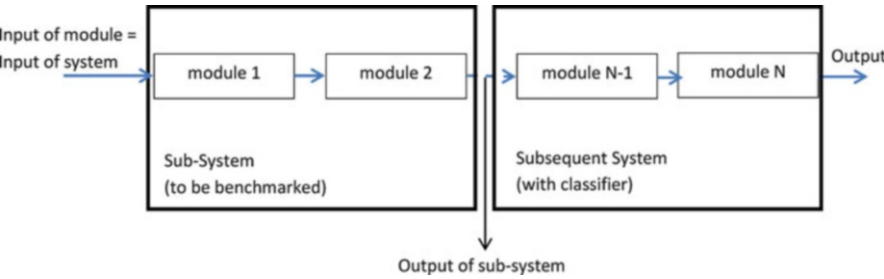
The most important result of this section is that GT depends on the kind of assessment function chosen. Table 30.2 presents an overview of ground-truthing tools developed for document analysis and processing tasks.

**Architecture of an Evaluation Framework**

In this section, the view of benchmarking is restricted to document analysis. Furthermore, a focus is set on the preprocessing steps of document analysis.

In benchmarking of document analysis systems, the same aims are pursued as for benchmarking in general. Not only the whole system is the subject of interest but also modules of this system are considered, because the improvement of modules of the system will cause an improvement of the overall system. Considering a document analysis system as a chain of modules, the detection of the weakest point in the chain might be the goal. This could not be done by looking at the system as a whole, thus a more detailed view is needed. In consequence, what is wanted to benchmark might be a single module, a sequence of modules or a subsystem of the overall system, see Fig. 30.9.

Benchmarking for whole systems had been conducted for years already. Generally accepted tests are conducted by institutions like NIST [13] and ISRI [27].



**Fig. 30.9** System in document analysis

In preprocessing, the need of benchmarks has been claimed but few have been published since today. This is mainly due to the fact that finding the ideal assessment is complicated, often subjective, and there is no standard describing test data sets with GT. These problems make benchmarking much harder in preprocessing than in other domains.

**The Assessment Function**

Benchmarking is possible when knowing GT. In order to know GT for real data, it has to be generated. A procedure to generate GT can be defined which is very similar to realizing a benchmark: the input data must be known, the subsystem or module, the assessment, and eventually the evaluation function. In most of the cases, these items will be identical to the ones for a benchmark. So, a focus is set on the generation of GT which includes all parts of the final benchmark. Since it is assumed that the input data and subsystem are given, finding the assessment function is the important task. Finding an ideal assessment function is nearly impossible, therefore one has to be satisfied with an approximation. In this context an objective validation of the assessment function is helpful.

Now, the focus is set first on the existent assessment methods and its references in literature. They may be defined by:

- A human
- The subsequent system or
- A direct comparison to ground truth

In the expert approach, a human evaluates the module’s output. The advantage of this method is its simplicity; however the disadvantage is that it is quite subjective. This approach was applied by Trier and Taxt [38] and by Nadal and Suen [26]. This method makes sense if the output of the system is produced at a machine/man interface or if the subsequent system is unknown.

The second approach is the replacement of the assessment function by the subsequent system up to a point where GT may easily be defined. For example, in Fig. 30.9 the modules 1 and 2 are the modules to benchmark. The following modules 3 till N represent the subsequent system, thus the assessment function. This is a widely used approach, e.g., Lam and Suen [18], Kanai, Rice, and Nartker [17],

Trier and Jain [37] and Lee, Park, and Tang [21] assessed a sub-system (pre-processing modules) using the results of a subsequent OCR system or other classifiers.

The third method is to replace the ideal assessment by a comparison with GT. In this case, GT has to be explicitly known. If synthetic data are used, GT is also known and a sufficient amount of data may be generated without difficulties. If real world data are used, GT can be defined by experts. The latter is close to the expert approach mentioned above. The shortcoming of real data and GT is the subjectiveness of GT and the difficulty to define rules to enable the human expert to generate objective GT. However, a test with real data produces the most meaningful results, whereas the synthetic data are always restricted in modeling the reality. Knowing the actual output and GT, a distance between them is computed to evaluate the quality of the actual output with respect to GT. (e.g., Haralick, Jaishima, and Dori [14], Lee, Lam, and Suen [20], Palmer, Dabis, and Kittler [28], Yanikoglu and Vincent [40], Randriamasy [32], and Randriamasy and Vincent [33] used this approach.)

These approaches inherit some problems. As the results are dependent on the assessment function and since the optimal assessment function can only be approximated, one has to be aware of errors and their consequences on the results. Performing a benchmark on an arbitrary interface, using the subsequent modules and their final result for assessment, poses the problem of not being able to identify clearly the origin of detected errors, which may lie either in the subsystem to be benchmarked or in the subsequent modules used in the assessment. The assumption of Kanai et al. [17] of an independence of these two types of errors may not hold. Furthermore, and even more important, when trying to generate GT directly at the benchmarking interface, one has to take into account that GT might not be clearly definable and will be faulty itself, making the identification of the error origin even more difficult. After presenting the existing approaches, it is important to point out that it is not evident to judge whether the system to benchmark or the assessment function is erroneous.

## Contests and Benchmarking

Testing recognition systems with large identical datasets is crucial for performance evaluation. Another challenge comes from their complexity because they consist of many specialized parts solving very diverse tasks. The recognition rate is a convenient measure for comparing different systems, but it is a global parameter hardly significant for system component development. To improve the overall system quality, it is essential to know the effectiveness of its modules.

The development of meaningful aspects of system evaluation methods was an important part of the aforementioned annual OCR tests at ISRI. The goal of these tests not only publicized the state of the art of page-reading systems but also provided information for improvement through competition and objective assessment. While much has been achieved concerning the evaluation problem (e.g., [34]),

**Table 30.3** Competitions and benchmarking campaigns overview

Competition/benchmarking	Conference	Topic(s)
<b>Pre-processing and document structure</b>		
Document image binarization contest	ICDAR 2009, 2011 and, ICFHR 2010	Image binarization
Page segmentation	ICDAR 2003, 2005, 2009	Page layout, structure
Text locating	ICDAR 2005	
Book structure extraction competition	ICDAR 2009, 2011	Book structure
Historical document layout analysis competition	ICDAR 2011	Layout analysis
Handwriting segmentation contest	ICDAR 2007, 2009 and ICFHR 2010	Segmentation
<b>Character/text recognition</b>		
OpenHaRT	2010, 2013	Word/text recognition
Arabic handwriting recognition competition	ICDAR 2005, 2007, 2009, 2011 and ICFHR 2010	Word/text recognition
Handwritten Farsi/Arabic character recognition competition	ICDAR 2009	Character recognition
French handwriting recognition competition	ICDAR 2009, 2011	Word/text recognition
Chinese handwriting recognition competition	ICDAR 2011	Word/text recognition
UNLV/ISRI evaluations	1992–1996	Annual tests of OCR accuracy
NIST evaluations	1994–1998	Word/text recognition
<b>Others</b>		
Signature verification competition	ICDAR 2009, 2011, and ICFHR 2010	Signature verification
Music scores competition: staff removal and writer identification	ICDAR 2011	Music score document analysis
Arc segmentation contest	GREC 2001, 2003, 2005, 2007, 2009	Arc segmentation

the availability of tools and data remains an issue for research, as discussed in the paper [25], published in 2005. For example, it is not enough to measure the quality, based on the symbol output of the recognizer, only by considering the word accuracy. The quality of zoning and the segmentation into words or characters represent an important feature of a recognition system and should be evaluated too [36]. A more general concept for evaluating system modules separately is presented in [24]. A list of competitions and benchmarking campaigns organized last 20 years are presented in Table 30.3. Databases, tools, and software used in different competitions and evaluations are collected and presented online by the IAPR TC10 and TC11 [16].

## Conclusion

This chapter describes tools and metrics useful for evaluation of document analysis systems or even parts of systems. To measure the performance of complex systems like systems for document analysis and recognition is a very difficult task. In early days first systems were evaluated only on private data sets; it followed that an objective quality measure was impossible. In recent years more and more data were made available to the research community which made it from that time possible to test and compare systems on same data sets. With these common data sets also common metrics were introduced to calculate a value of system quality. The most challenging task is the evaluation of segmentation modules as in this case not only symbols but also segments have to be compared in size and position. Metrics for different modules of the system are discussed in this chapter; some of them are consolidated and used frequently. Increasing interest in document system evaluation also becomes apparent in more and more open competitions for different fields of applications. Finally tools for ground-truth generation and system benchmarking are presented.

Most interesting new perspective in system quality measure is the fact that tools to support a fast and semiautomatic ground-truth generation allow an effective training with adaptation to a given data distribution. Together with a detailed quality measure even of parts of a system, only a fast system optimization is possible and a better understanding of the weakness of system parts.

Anybody interested in more details is recommended to read papers presenting the results of the competitions or system benchmarks.

---

## References

1. Antonacopoulos A, Gatos B, Karatzas D (2003) ICDAR 2003 page segmentation competition. In: Proceedings the 7th international conference on document analysis and recognition (ICDAR), Edinburgh, pp 688–692
2. Antonacopoulos A, Bridson D, Gatos B (2005) Page segmentation competition. In: Proceedings the 8th international conference on document analysis and recognition (ICDAR), Seoul, pp 75–79
3. Antonacopoulos A, Karatzas D, Bridson D (2006) Ground truth for layout analysis performance evaluation. In: Proceedings of the 7th international conference on document analysis systems (DAS 06), Nelson, pp 302–311
4. Antonacopoulos A, Gatos B, Bridson D (2007) Page segmentation competition. In: Proceedings the 8th international conference on document analysis and recognition (ICDAR), Curitiba, pp 1279–1283
5. Baird HS, Govindaraju V, Lopresti DP (2004) Document analysis systems for digital libraries: challenges and opportunities. In: Proceedings of the IAPR international workshop on document analysis systems (DAS), Florence, pp 1–16
6. Belaid A, D'Andecy VP, Hamza H, Belaid Y (2011) Administrative document analysis and structure. In: Learning structure and schemas from documents. Springer, Berlin/Heidelberg, pp 51–72



7. Bippus R, Märgner V (1995) Data structures and tools for document database generation: an experimental system. In: Proceedings of the third international conference on document analysis and recognition (ICDAR 1995), Montreal, 14–16 Aug 1995, pp 711–714
8. Chhabra A, Phillips I (1998) The second international graphics recognition contest – raster to vector conversion: a report. In: Graphics recognition: algorithms and systems. Lecture notes in computer science, vol 1389. Springer, Berlin/Heidelberg, pp 390–410
9. Clausner C, Pletschacher S, Antonacopoulos A (2011) Aletheia – an advanced document layout and text ground-truthing system for production environments. In: Proceedings of the 11th international conference on document analysis and recognition (ICDAR), Beijing, pp 48–52
10. Doermann D, Zotkina E, Li H (2010) GEDI – a groundtruthing environment for document images. In: Proceedings of the ninth IAPR international workshop on document analysis systems, DAS 2010, Boston, 9–11 June 2010
11. Fenton R (1996) Performance assessment system development. *Alsk Educ Res J* 2(1):13–22
12. Gatos B, Ntirogiannis K, Pratikakis I (2011) DIBCO 2009: document image binarization contest. *Int J Doc Anal Recognit* (Special issue on performance evaluation) 14(1):35–44
13. Geist J et al (ed) (1994) The second census optical character recognition systems conference. Technical report NISTIR-5452, National institute of standards and technology, U.S. Department of Commerce, U.S. Bureau of the Census and NIST, Gaithersburg
14. Haralick RM, Jaisimha MY, Dori D (1993) A methodology for the characterisation of the performance of thinning algorithms. In: Proceedings of the ICDAR'93, Tsukuba Science City, pp 282–286
15. Héroux P, Barbu E, Adam S, Trupin É (2007) Automatic ground-truth generation for document image analysis and understanding. In: Proceedings of the 9th international conference on document analysis and recognition (ICDAR 2007), Curitiba, Sept 2007, pp 476–480
16. IAPR TC11 Website for Datasets, Software and Tools (2012). [http://www.iapr-tc11.org/mediawiki/index.php/Datasets\\_List](http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List), Dec 2012
17. Kanai J, Rice V, Nartker TA (1995) Automated evaluation of OCR zoning. *IEEE Trans TPAMI* 17(1):86–90
18. Lam L, Suen CY (1993) Evaluation of thinning algorithms from an OCR viewpoint. In: Proceedings of the ICDAR'93, Tsukuba Science City, pp 287–290
19. Lee CH, Kanungo T (2003) The architecture of TRUEVIZ: a groundtruth/metadata editing and visualizing toolkit. *Pattern Recognit* 36(3):811–825
20. Lee S-W, Lam L, Suen CY (1991) Performance evaluation of skeletonization algorithms for document analysis processing. In: Proceedings of the ICDAR'91, Saint Malo, pp 260–271
21. Lee S-W, Park J-S, Tang YY (1993) Performance evaluation of nonlinear shape normalization methods for the recognition of large-set handwritten characters. In: Proceedings of the ICDAR'93, Tsukuba Science City, pp 402–407
22. Lu H, Kot AC, Shi YQ (2004) Distance-Reciprocal distortion measure for binary document images. *IEEE Signal Process Lett* 11(2):228–231
23. Lucas SM (2005) Text locating competition results. In: Proceedings of the 8th international conference on document analysis and recognition (ICDAR), Seoul, pp 80–85
24. Märgner V, Karcher P, Pawlowski AK (1997) On benchmarking of document analysis systems. In: Proceedings of the 4th international conference on document analysis and recognition (ICDAR), Ulm, vol 1, pp 331–336
25. Märgner V, Pechwitz M, El Abed H (2005) ICDAR 2005 Arabic handwriting recognition competition. In: Proceedings of the 8th international conference on document analysis and recognition (ICDAR), Seoul, vol 1, pp 70–74
26. Nadal C, Suen CY (1993) Applying human knowledge to improve machine recognition of confusing handwritten numerals. *Pattern Recognition* 26(3):381–389
27. Nartker TA, Rice SV (1994) OCR accuracy: UMLVs third annual test. *INFORM* 8(8):30–36
28. Palmer PL, Dabis H, Kittler J (1996) A performance measure for boundary detection algorithm. *Comput Vis Image Underst* 63(3):476–494

29. Phillips I, Chhabra A (1999) Empirical performance evaluation of graphics recognition systems. *IEEE Trans Pattern Anal Mach Intell* 21(9):849–870
30. Phillips I, Liang J, Chhabra A, Haralick R (1998) A performance evaluation protocol for graphics recognition systems. In: *Graphics recognition: algorithms and systems. Lecture notes in computer science*, vol 1389. Springer, Berlin/Heidelberg, pp 372–389
31. Pratikakis I, Gatos B, Ntirogiannis K (2011) ICDAR 2011 Document image binarization contest (DIBCO 2011). In: *Proceedings of the 11th international conference on document analysis and recognition*, Beijing, Sept 2011, pp 1506–1510
32. Randriamasy S (1995) A set-based benchmarking method for address bloc location on arbitrarily complex grey level images. In: *Proceedings of the ICDAR'95*, Montreal, pp 619–622
33. Randriamasy S, Vincent L (1994) Benchmarking and page segmentation algorithms. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'94)*, Seattle, pp 411–416
34. Rice SV (1996) Measuring the accuracy of page-reading systems. PhD thesis, Department of Computer Science, University of Nevada, Las Vegas
35. Saund E, Lin J, Sarkar P (2009) PixLabeler: user interface for pixellevel labeling of elements in document images. In: *Proceedings of the 10th international conference on document analysis and recognition (ICDAR2009)*, Barcelona, 26–29 July 2009, pp 446–450
36. Thulke M, Märgner V, Dengel A (1999) A general approach to quality evaluation of document segmentation results. In: *Document Analysis Systems: Theory and Practice. Third IAPR workshop DAS'98*, Nagano, Japan, selected papers. LNCS vol 1655. Springer, Berlin/Heidelberg, pp 43–57
37. Trier OD, Jain AK (1995) Goal-directed evaluation of binarization methods. *IEEE Trans PAMI* 17(12):1191–1201
38. Trier OD, Taxt T (1995) Evaluation of binarization methods for document images. *IEEE Trans PAMI* 17(3):312–315
39. Yacoub S, Saxena V, Sami SN (2005) PerfectDoc: a ground truthing environment for complex documents. In: *Proceedings of the 2005 eight international conference on document analysis and recognition (ICDAR'05)*, Seoul, pp 452–457
40. Yanikoglu BA, Vincent L (1998) Pink Panther: a complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognit* 31(9):1191–1204
41. Yanikoglu BA, Vincent L (1995) Ground-truthing and benchmarking document page segmentation. In: *Proceedings of the third international conference on document analysis and recognition (ICDAR)*, Montreal, vol 2, pp 601–604

## Further Reading

Many books about performance evaluation and benchmarking are on the market, especially for benchmarking of computer systems. But there is no book about document analysis methods evaluation. In most image processing, pattern recognition, and document analysis books chapters about evaluation of methods can be found. Readers interested in modern evaluation and benchmarking methods in general may find more details in the following recently published books:

Madhavan R, Tunstel E, Messina E (eds) (2009) *Performance evaluation and benchmarking of intelligent systems*. Springer, New York

Obaidat MS, Boudriga NA (2010) *Fundamentals of performance evaluation of computer and telecommunication systems*. Wiley, Hoboken