Introduction to Document Analysis and Recognition

Simone Marinai

University of Florence
Dipartimento di Sistemi e Informatica (DSI)
Via S. Marta, 3, I-50139, Firenze, Italy
marinai@dsi.unifi.it

Document Analysis and Recognition (DAR) aims at the automatic extraction of information presented on paper and initially addressed to human comprehension. The desired output of DAR systems is usually in a suitable symbolic representation that can subsequently be processed by computers.

Over the centuries, paper documents have been the principal instrument to make permanent the progress of the humankind. Nowadays, most information is still recorded, stored, and distributed in paper format. The widespread use of computers for document editing, with the introduction of PCs and word-processors in the late 1980's, had the effect of increasing, instead of reducing, the amount of information held on paper. Even if current technological trends seem to move towards a paperless world, some studies demonstrated that the use of paper as a media for information exchange is still increasing [1]. Moreover, there are still application domains where the paper persists to be the preferred media [2].

The most widely known applications of DAR are related to the processing of office documents (such as invoices, bank documents, business letters, and checks) and to the automatic mail sorting. With the current availability of inexpensive high-resolution scanning devices, combined with powerful computers, state-of-the-art OCR packages can solve simple recognition tasks for most users. Recent research directions are widening the use of the DAR techniques, significant examples are the processing of ancient/historical documents in digital libraries, the information extraction from "digital born" documents, such as PDF and HTML, and the analysis of natural images (acquired with mobile phones and digital cameras) containing textual information.

The development of a DAR system requires the integration of several competences in computer science, among the others: image processing, pattern recognition, natural language processing, artificial intelligence, and database systems. DAR applications are particularly suitable for the incorporation of

S. Marinai: *Introduction to Document Analysis and Recognition*, Studies in Computational Intelligence (SCI) **90**, 1–20 (2008)
www.springerlink.com

machine learning techniques for two factors: first, classification algorithms are used at several processing levels, from image pre-processing to character classification; second, large collections of manually annotated document images are available and can be used for automatic training of classifiers. As a matter of fact, in the last decades isolated handwritten character recognition has been frequently used as a standard benchmark for evaluating and comparing machine learning algorithms. However, besides isolated character recognition there are several other sub-tasks that have been tackled with machine learning techniques.

In spite of these considerations, there are several systems described in the DAR literature that address relevant sub-tasks with manually tuned algorithms without resorting to machine learning techniques. The aim of this book is to link together the DAR research with the machine learning one. The chapters in the book cover the state of the art in several DAR sub-tasks and include also inspiring pointers to future research directions. The DAR literature is large and covered by several books and survey papers [3, 4, 5, 6, 7, 8]. In this introductory chapter, we provide a brief guide to the DAR field providing pointers to the relevant literature and to benchmark databases. We propose also an overview of the contents of this book: the remaining fifteen chapters cover many sub-tasks in DAR applications providing some insights in the current research trends.

1 DAR Applications

There are many examples of the use of DAR techniques into both commercial and research-driven systems. Some systems have been in use (with continuous improvements) for decades and are now widely identified as successful DAR applications. We can split the DAR applications into two broad categories: business-oriented and user-centered ones.

Office documents reach a total of more than 85% of the amount of new original information stored on paper in the world [1]. It is therefore not surprising that business-oriented applications received a great interest. In this category we can include the automatic check processing (including both amount reading and signature verification), the information extraction from forms and invoices, the automatic document organization that involves page classification.

One well known application of handwriting recognition is the automatic postal mail sorting based on address recognition on envelopes. Two main problems are addressed in these systems. First, the layout of the envelope image is processed so as to identify the address position. Second, the address is recognized, avoiding confusion between sender and recipient.

Among the user-centered applications we include the software tools, such as OCR software for general purpose PCs, that can be used to process personal information originated in paper form. Other applications that received

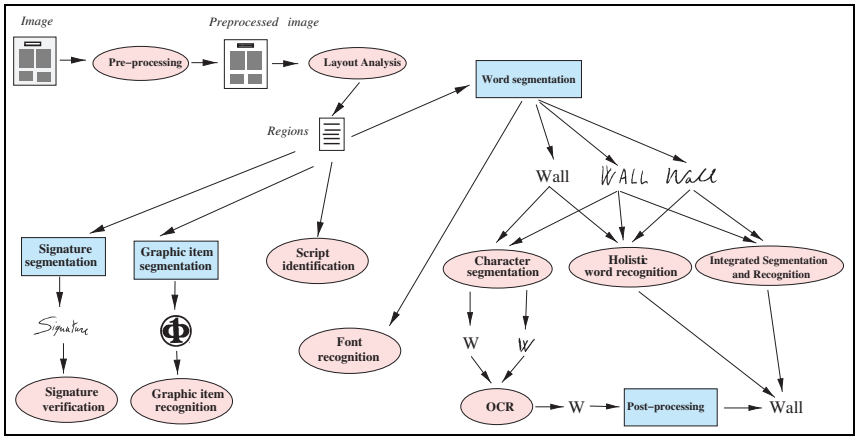

Fig. 1. General view of the document processing data-flow. Oval boxes correspond to tasks approached with significant results using machine learning-based methods. In the right part, printed, hand-printed, and handwritten instances of the word "Wall" are taken as examples to show different processing flows related to different writing sources

attention in recent years are related to the development of tools aimed at improving the access to the objects in digital libraries and processing of historical documents. In this context, large collections of digitized documents are now available in Internet, to both scholars and the general public. The information extraction from these documents (mostly book and journals) allows users to retrieve information related to the cultural heritage, as well as to identify novels and essays of interest. Further applications in this category are the recognition of printed music score and the analysis of drawings such as maps and cadastral maps. Recently, portable devices, such as mobile phones and PDAs, has been considered to provide input on-the-fly for camera-based document processing systems[1].

In most DAR applications the document content is conceptually described by means of the *physical* and the *logical* structures. The *physical* structure describes the visual aspect of the document by representing the basic objects and their mutual positions. The *logical* structure assigns to each object a suitable meaning.

2 Processing Steps in DAR

A complex DAR system is organized similarly to most Pattern Recognition (PR) systems including four principal components: pre-processing, object

[1] Some recent references can be found in the *Camera Based Document Analysis and Recognition* (CBDAR) workshop proceedings: http://www.m.cs.osakafu-u.ac.jp/cbdar2005

segmentation, object recognition, and post-processing. The *pre-processing* aims at improving the quality of the images. The *object segmentation* allows to identify the basic objects in the document. In DAR this task takes different names, depending on the application level considered. When dealing with regions in the page the segmentation is referred to as layout analysis (see Section 4). At a lover processing level we deal with various objects such as signature, word, character, and so on. The *object recognition*, or classification, deals with the objects identified in the previous step. We analyze in Section 6 the various levels at which object classification can occur in DAR applications. Lastly, the *post-processing* checks the results of the classification on the basis of contextual information.

To describe with more details the peculiarities of DAR systems we depict in Fig. 1 the relationships between the main document processing tasks. *Pre-processing* operations in DAR are used in order to improve the input image for subsequent analysis. Common tasks are de-skew, image enhancement (noise reduction) and character thinning. *Layout analysis* methods are aimed at extracting the physical and/or logical structure of the document image. When dealing with the textual parts of a document, the words are usually located by *word segmentation* with low-level methods such as morphological processing and connected components clustering. It is important to remark that in some languages, like Japanese and Chinese, there is no word separation by spaces as in western languages such as English. Word segmentation is required for the recognition of printed, hand-printed, and cursive text. The subsequent word recognition can be based either on the segmentation and subsequent recognition of the individual characters, or on the recognition of the whole word image as a single unit. In the former approach, based on a "divide-et-impera" scheme, a preliminary step of *character segmentation* is required and the isolated characters are subsequently recognized. In *holistic word recognition* the whole word is considered as a single object to be recognized. In so doing errors due to wrong character segmentations can be avoided. However, this approach can be effectively used only in the presence of a limited dictionary for instance in postal applications and check reading. An alternative approach is based on *integrated segmentation and recognition*, where the two operations take place at the same time. *Graphical items* are recognized with methods similar to those applied to character recognition, whereas specific approaches are considered for dealing with *signature verification*.

Important DAR tasks are also *script identification* that helps processing multi-lingual languages, and *font recognition* that can contribute to improve the performance of text reading. In addition to text-intensive documents, a large group of DAR applications deals with graphical documents such as technical drawings and maps, where different recognition methods are adopted.

Similarly to other PR applications there are two principal approaches to document analysis: *top-down* and *botttom-up*. If the broad structure of the document layout is known in advance, then a model-driven (top-down) approach can be used. When the physical structure is not known in advance,

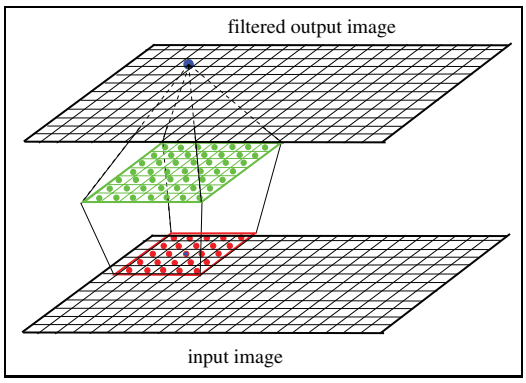

Fig. 2. An MLP acting as a filter. The input is transformed onto the output by moving the input window in raster order

then a data-driven (bottom-up) approach should be considered. In the latter case all the objects in the image must be located and recognized in order to identify the desired information. The peculiarities of top-down and bottom-up approaches in layout analysis are discussed in particular in the chapter by Belaïd and Rangoni.

In the following sections we provide a deeper analysis of the most important tasks in DAR with pointers to appropriate chapters in this book.

3 Pre-processing

The document acquisition is the process of obtaining an electronic image of a paper-based document. In most cases a flat-bed scanner is used, however in digital libraries also book scanners can be considered, whereas in recent years portable devices such as digital cameras and mobile phones are used as well.

Pre-processing operations in document image analysis transform the input image into an enhanced image more suitable for further analysis. Image-to-image transformations in DAR belong to four main classes [9]: filtering, geometrical transformations (e.g. skew detection), object boundary detection, and thinning.

The filtering transforms the input into an image whose value in a generic position (i, j) is usually a function of the input values in a neighborhood of the point (i, j). The three main classes of filtering operations in DAR ([4], page 3) are *binarization*, *noise reduction*, and *signal enhancement*. An example of machine-learning applications for filtering is based on the well-known property of universal approximation of neural networks, and in particular of the Multi-Layer Perceptron (MLP) [10]. To this purpose, a trained MLP (or other supervised classifiers) is fed with the pixels of a fixed-size sliding window providing a suitable output image (Fig. 2).

Thinning algorithms are needed to compute features based on the symbol skeleton. The simplest approaches are based on a recursive erosion of the object contour (for instance by using morphological operations). Other approaches rely on clustering-based skeletonization algorithm (CBSA) that is also discussed in the chapter by Marinai, Marino, and Soda.

4 Layout Analysis

Printed pages contain various types of information (e.g. text, equations, graphics, images) that should be processed with suitable techniques. Document layout analysis is performed to segment the document image into regions having an homogeneous content and to assign a meaning to the regions (Fig. 3). The segmentation step is called *physical layout analysis*, and is used to identify the geometric page structure. The *logical layout analysis*, or functional labeling, assigns a logical meaning to each region generating the logical structure of the document. Further remarks on these topics can be found in the chapter by Malerba, Ceci, and Berardi. A layout analysis sub-task that is particularly difficult to tackle is table recognition that is addressed in this book in the chapter by Zanibbi, Blostein and Cordy.

4.1 Physical Layout Analysis

The physical layout analysis aims at extracting regions with uniform features from the document. Segmentation algorithms in image processing can be grouped into three main categories: pixel classification, edge-based segmentation, and region-based segmentation. More specifically, document segmentation approaches pertain to pixel classification and region-based segmentation.

In pixel classification a label is attached to each pixel considering its color and those of the neighboring pixels. It was initially applied to the binarization of document images (e.g. [11]) and then extended to deal with additional classes (e.g. text, graphics, and line drawing). The regions are subsequently extracted by removing small noisy elements, merging similar neighboring regions, and locating connected components in the resulting image.

Most segmentation methods in DAR belong to the region-based family, comprising bottom-up and top-down methods. Bottom-up approaches can be regarded as merging methods and are mostly based on the location of connected components, with subsequent aggregation in higher level structures. In the top-down analysis a page is segmented from larger components to smaller subcomponents (e.g. [12, 13]). For instance, a page can be split into columns, in paragraphs, text lines, words and characters. Several methods (e.g. the XY tree) are based on the computation of the projection profile by computing the number of black pixels accumulated in horizontal and vertical directions to identify gaps between regions in the page. Top-down methods are usually faster but work well only for documents having a regular layout.

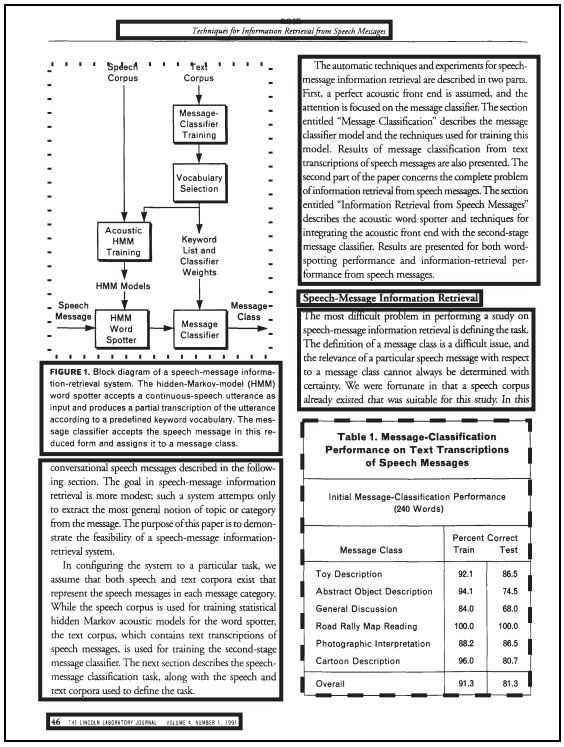

Fig. 3. Example of layout segmentation. Text blocks are identified with continuous lines. A dotted line and a dashed one identify one table region and one picture region, respectively

Bottom-up methods are more expensive from a computational point of view, but can deal with more complex layouts.

4.2 Logical Layout Analysis

The logical layout analysis assigns a meaning to the regions identified by the physical layout analysis. Examples of features considered are the size of the blocks, their mutual position, and some information on the textual parts such as the predominant font, character size and spacing. In this book the use of the logical classification of textual regions in a digital library context is described in the chapter by Esposito, Ferilli, Basile, and Di Mauro.

The physical and logical analysis can be performed together [14] so as to assign a meaning to blocks during their segmentation. In most cases this is not feasible, since the class can be defined only after analyzing the region position with respect to other parts of the page (or even after the reading of its content).

5 Text Recognition

The text recognition is executed to convert an image containing printed or handwritten text into a format that can be understood by a computer (e.g. ASCII or unicode). State-of -the-art OCR packages can read printed text with very high recognition rates. However, the research on printed text recognition is still in progress in order to deal with noisy documents and non-standard fonts (see e.g. the chapter by Marinai, Marino, and Soda), as well as large character sets (see e.g. the chapter by Jaeger, Ma, and Doermann and the survey chapter by Liu and Fujisawa).

Text reading techniques can be split into two main categories (*on-line* and *off-line*) on the basis of the input device used. The on-line recognition uses tablet PCs as input and the text is acquired and processed in real time when written by the user, whereas scanners are often used in off-line systems. One significant difference between on-line and off-line systems is the nature of the signal considered as input to the recognizer. In the on-line recognition the signal is dynamic and represents the text drawn by the user with one sequence of values that describes the pen position and pressure information. In off-line processing the input signal is a raster image. The position of the text in the page is usually unknown and should be identified by layout analysis algorithms.

In on-line systems there are some advantages with respect to off-line processing since the temporal information of the writing is known. This information is helpful to segment individual characters and to identify character's strokes. A more subtle advantage of on-line systems is the interactivity that is established between the user and the system. This interaction helps the automatic training of reading systems, but is mostly important for user adaptation to the system (related issues of user interaction are addressed in the chapter by Nagy and Veeranachaneni). The combination of on-line and off-line recognition into a multiple classifier framework is proposed in the chapter by Jaeger, Ma, and Doermann.

Some salient features of off-line reading systems are discussed in the rest of this section.

5.1 Character Recognition

In the *divide-et-impera* paradigm, that is typical of several pattern recognition applications, the text recognition systems frequently first split the words into characters and then assign one class to each isolated object. When dealing with printed text this approach is generally referred to as Optical Character Recognition (OCR) that is exploited by several software tools available on the market.

Summarizing, the character recognition process can be divided into three main steps: 1) *segmentation and preprocessing*, are used to identify the characters and improve the image quality; 2) *classification* is typically based on a

sequential application of a feature extraction module and a supervised classifier; 3) *contextual processing* is adopted to check the recognition results on the basis of contextual information such as domain-specific dictionaries.

The first patent of a system for the automatic recognition of printed characters was registered in 1929 [3]. The system used electro-optical techniques to recognize characters. There is a large literature on character recognition and OCR with many technical papers dating back to the 1960's. We do not have the ambition of discussing in detail this literature. Instead, we point out a few survey papers [3, 6, 8] and the chapter by Liu and Fujisawa in this book that reviews the learning-based classification methods that have been applied to character recognition.

5.2 Word Recognition

The segmentation-based word recognition cannot be exploited when the location of segmentation points is impossible or unreliable, as in cursive handwriting. In this case one alternative is to use a holistic recognition where words are recognized as single units. Holistic recognition is effective when dealing with a reduced lexicon. For instance, the number of basic words required to fill a check is limited to a few dozen (32 words are needed for writing legal amounts of English checks [15], 30 words can be used for French checks [16]).

When the problem at hand requires a larger lexicon, then segmentation-based reading is appropriate, nevertheless segmentation requires some feedback from recognition. Integrated segmentation and recognition (ISR) techniques are related to the contextual development of segmentation and recognition modules. Two well known ISR techniques are Heuristic Over Segmentation (HOS) and Time-Delay Neural Networks (TDNN).

In HOS a segmentation algorithm is applied to a word image to locate a large number of candidate cutting points. Subsequently, a recognizer is employed to score the alternative segmentations generated and to find the best character sequence. The basic idea behind this approach is to over-segment the word in the hope of including all the correct segmentation points among those extracted. The use of MLP for labeling valid segmentation points is described in the chapter by Blumenstein, whereas a related application in the field of postal address reading is discussed in the chapter by Kagehiro and Fujisawa. In the chapter by Tulyakov and Govindaraju the combination of handwritten word recognizers based on the HOS approach is analyzed.

Time-Delay Neural Networks (TDNN) have been initially proposed to deal with temporal sequences. The output of a TDNN depends on its current and previous inputs, which are delayed by one or more time units. If the input signal is a vector of m values and we take a delay of n time units into account, then a TDNN can be implemented with an MLP having $n \cdot m$ input units ([17] page 256). TDNNs have been used for on-line and off-line word recognition. In on-line recognition the meaning of "time" is quite straightforward, whereas in the off-line case the horizontal axis in the window

containing the word is considered as a temporal scale. In this book the chapter by Belaïd and Rangoni contains a deep analysis of TDNN.

6 Classification in DAR

Classifiers trained by supervised learning are key components of many pattern recognition systems and DAR systems are not exceptions. In this section we summarize the various DAR sub-tasks that have been addressed with supervised classifiers. Another important family of methods is based on unsupervised classifiers whose applications in DAR are discussed in the chapters by Nagy and Veeramachaneni and by Marinai, Marino, and Soda. Since the book is at the cross-road of document analysis and machine learning, it is not surprising that many chapters deal with the topics summarized below.

6.1 Pixel and Region Classification

Pixel classification (Section 4.1) can be used for both pre-processing and layout analysis. Document image binarization is a simple pre-processing task that can be performed with pixel classification. The assumption behind most pixel classification approaches is that textual and graphical regions have different textures, and thus the membership of each pixel can be estimated by analyzing a small region around it (e.g. [18]).

Region classification is used in most cases in conjunction with global features describing the whole region to be labeled. The classification was initially carried out using linear classifiers operating on these features according to user-defined parameters [19]. The basic hypothesis that is exploited in region classification is the assumption that the content of the region is homogeneous, and consequently some general features of the whole region can be used as inputs to a trainable classifier (e.g. [20]). The output of the classifier is a logic role that can be associated to each region.

6.2 Reading Order Detection

The task of text reading cannot be limited to the simple recognition of the individual characters or words. Instead, the identification of the correct reading order allows human readers to correctly understand the document content. The identification of the correct word sequence can be extremely complex when dealing with multi-column documents with footnotes and when figure and tables are intermixed in the text. The chapter by Malerba, Ceci, and Berardi formulates the reading order problem representing reading order chains in first order logic formalism. In the proposed system the learned rules state that one block "follows in reading" another block on the basis of the block position in the page and other textual features.

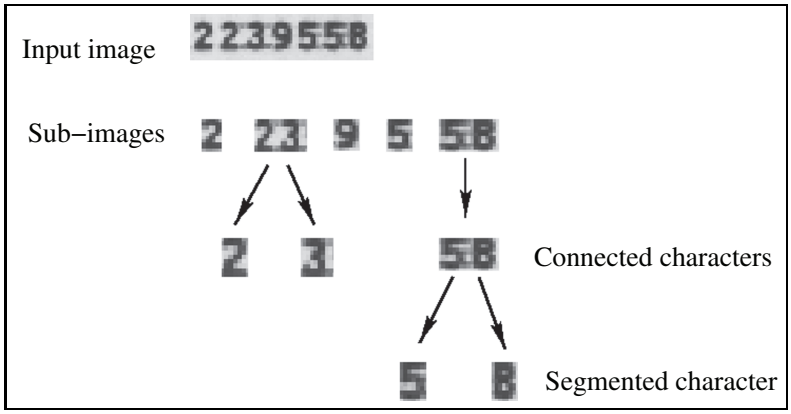

Fig. 4. Segmentation of connected strings. Supervised classifiers can be used for both identifying and segmenting the connected characters (as in '5B')

6.3 Text Recognition

In Section 5 we summarized the main features of text recognition systems. Various chapters in this book address the problem of word and character recognition. Liu and Fujisawa provide an updated survey of the use of learning methods in handwritten character recognition. Varga and Bunke propose the use of synthetic training data to improve the performance of an off-line cursive handwriting recognition system. One important research topic is the study of suitable techniques for combining classifiers so as to improve the overall classification performance. The chapter by Tulyakov, Jaeger, Govindaraju, and Doermann provides an overview of the classifier combination methods that have been addressed in several DAR tasks and in particular in character and word recognition.

Strictly related with word recognition is word retrieval that aims at efficiently retrieving occurrences of user-defined query words. The application of unsupervised learning to the task of word retrieval is discussed in the chapter by Marinai, Marino, and Soda.

6.4 Character segmentation

Supervised classifiers can be used in character segmentation with two main goals: the identification of touching characters and the location of cutting points (Fig. 4). In the former case the classifier is trained to distinguish isolated characters from pairs of touching characters (e.g. [21]) or more generally to estimate the number of characters in a string of connected digits (e.g. [22]). In the latter case the classifier is used for locating cutting points, i.e. the position in the word image that divides touching characters. Similarly to neural

filters a classifier can be horizontally "moved" across the input image with the aim of labeling the corresponding position.

When dealing with handwritten strings the segmentation points can be located by analyzing the primitives (horizontal strokes) instead of the raw image. This approach is considered in [23] where an MLP is used in order to identify strokes that can correspond to cutting points. In this book the chapter by Blumenstein proposes a neural-based technique for classifying segmentation points in a cursive character segmentation algorithm.

6.5 Script Identification

Multi-lingual documents are important in particular in countries, like India, where there are multiple scripts that can be mixed together in the same text [24]. The script identification is a preliminary stage, with respect to the language recognition, that aims at recognizing the script used in a fragment of text. In some applications the identification is addressed at the page level. In such cases texture-based algorithms can be considered. In other applications there is more interest in the identification of the script of single words. In the chapter by Jaeger, Ma, and Doermann the combination of classifiers with informational confidence is tested, among other tasks, to the script identification in bilingual (e.g., Arabic-English) dictionaries.

6.6 Signature Verification

The task of signature verification systems is to evaluate whether unknow signatures are genuines or forgeries. At first one can imagine to solve the problem with a classifier trained to distinguish among two classes: genuine and forgery. However, several problems arise when trying to implement actual systems. For instance, we should deal both with random forgeries (affixed without knowing the actual signature) and with skilled ones (where the right signature is known to the forger). These issues are discussed in the chapter by Srihari, Srinivasan, Chen, and Beal that addresses the use of machine learning techniques for signature verification.

6.7 Writer Identification

Together with signature verification, writer identification, that is aimed at identifying the author of an handwritten text, is one of the oldest biometrics used for forensic purposes. In addition to these traditional applications, writer identification can be suitably incorporated in biometrics technology that is being adopted in various security applications. Writer identification systems can be either text-independent or text dependent. The former approach is the most difficult, since any text can be used to establish the identity of a writer. In this book the chapter by Schlapbach and Bunke presents an off-line, text-independent system for writer identification and verification.

6.8 Page Classification

In page classification the input to the classifier is the whole page. This task is useful when processing large collections of documents, for instance to perform an automatic document indexing. Page classification methods can be used in many application domains, such as form processing systems and digital libraries. Earlier applications concerned form classification methods that are aimed at selecting an appropriate reading strategy for each form to be processed. These methods often take the presence of ruling lines in the preprinted form layout into account [25]. Other typical classes are business letters and technical papers [26]. In the last few years the classification of journal and book pages received greater attention for digital library applications [27, 28].

Page classification in the domain of scientific documents is discussed in the chapter by Esposito, Ferilli, Basile, and Di Mauro where multi-page documents are indexed by matching their first page against some automatically learned models of document classes. Page classification for invoice processing is addressed in the chapter by Marinai, Marino, and Soda.

6.9 Document Categorization

Page classification has some similarities with document categorization, where documents are classified on the basis of the topic they address. Document categorization can be obtained by taking into account the document textual content (possibly recognized with an OCR software). The application of document categorization techniques to PDF documents is discussed in the chapter by Esposito, Ferilli, Basile, and Di Mauro. In some application domains it is possible to attempt the document categorization relying on the word image representation [29]. The latter strategy is addressed in the chapter by Marinai, Marino, and Soda.

7 Training Data

The tuning and comparison of machine learning algorithms heavily relies on the use of large collections of annotated data to be used to train the algorithms. In particular, the recognition of isolated handwritten digits has been considered for a long time as a benchmark for the comparison of trainable classifiers. For example the dataset "Optical Recognition of Handwritten Digits" is hosted since 1998 in the well known *UCI Machine Learning Repository* [2]. We survey in this section some of the most widely used public domain databases in the DAR field. We attempted to mention the most important datasets (including those collected to perform competitions at main conferences), without aiming to be exhaustive. Digital libraries are another important source of data

[2] http://mlearn.ics.uci.edu/MLSummary.html

for research in DAR. Some researchers built databases starting from widely available data such as free DLs or web sites of main publishers (e.g. IEEE). Examples of these datasets are adopted in the chapters by Esposito et al., and Marinai et al.

7.1 Public Databases

The databases listed in this section have been used by different research groups to compare the performance of DAR systems addressing several tasks. In most cases the data are available on the Internet and can be freely downloaded for research purposes.

UW databases

The UW databases are three document collections that have been gathered and manually annotated by the Intelligent Systems Laboratory, at the University of Washington (WA) in the late 1990's under the supervision of Prof. Robert Haralick. The databases were distributed as CD-ROMs containing document images and software for research and development.

The UW-III is the third in the series (it was published in 1996) and contains pages of chemical and mathematical equations, pages of line drawings and engineering drawings. There are also 33 pages containing English text with the corresponding character level groundtruth, 979 pages from UW-I, and 623 pages from UW-II corrected for skew, and the word bounding boxes for each word on these pages. These CD-ROMs, distributed for research purposes under the payment of a small fee, have been a reference for many years for the research on printed text.

NIST databases

In the 1990's the National Institute of Standards and Technologies (NIST) produced several CD-ROMs aimed at supporting the research on OCR software and information retrieval systems. A fully-automated process developed at NIST was used to derive the groundtruth information for the document images. The method involves matching the OCR results from a page with typesetting files for an entire book. The databases included scanned images, SGML-tagged groundtruth text, commercial OCR results, and image quality assessment results.

NIST's SD-3 (Special Database 3) and SD-1 contained binary images of handwritten digits. NIST originally designated SD-3 as training set and SD-1 as test set. However, SD-3 is much cleaner and easier to recognize than SD-1 and this fact is a limit for comparing different algorithms.

MNIST dataset

To overcome the previously mentioned limits of the NIST datasets Yann LeCun [30] designed the MNIST database of handwritten digits with a training set of 60,000 examples, and a test set of 10,000 examples[3]. MNIST is a subset of a larger collection of data available from NIST. The digits have been size-normalized and centered in a fixed-size image. The database has been widely used in the last few years for comparing various classifiers on real-world data [30, 31].

MediaTeam database

The MediaTeam Oulu Document Database[4] is a collection of 500 scanned document images with related groundtruth information about the physical and logical structure of the documents. The images cover a broad range of document types including journal papers, maps, newsletters, form, music, dictionaries and can be used for comparing various tasks in DAR.

Infty project

The InftyProject is a voluntary R&D organization consisting of researchers from different universities and research institutes in Japan with the shared objective of investigating and developing new systems to process scientific information by computer. Starting from 2005 three datasets have been distributed[5]. InftyCDB-1 [32] consists of mathematical articles in English containing 688,580 objects (characters and mathematical symbols) from 476 pages. The image of each object is recorded together with appropriate groundtruth information. InftyCDB-2 has the same structure of InftyCDB-1 and contains some documents in German and French, as well as many in English. InftyCDB-3 is a database of single alphanumeric characters and mathematical symbols. Unlike InftyCDB-1 and InftyCDB-2, word and mathematical expression structure is not included. The images are of individual characters only for a total of 259,389 symbols.

IFN/ENIT database

The IFN/ENIT-database[6] is based on more than 2,200 Arabic handwritten forms (filled by Tunisian people) containing about 26,000 word images together with groundtruth information [33]. This database has been used as a basis for a competition on Arabic handwriting recognition that is described below.

[3] http://yann.lecun.com/exdb/mnist
[4] http://www.mediateam.oulu.fi/downloads/MTDB
[5] http://www.inftyproject.org/en/database.html
[6] http://www.ifnenit.com

MARG database

MARG[7] is a freely-available repository of document page images and their associated groundtruth information on the textual and layout content. The pages are representative of articles drawn from the corpus of important biomedical journals. The database is suitable for the development of algorithms to locate and extract text from the bibliographic fields typical of articles within such journals. These fields include the article title, author names, institutional affiliations, abstracts and possibly others. Only the first page of each article is available, or the second page if the abstract runs over [34].

IAM database

The IAM-Database[8] includes over 1,500 scanned forms of handwritten text from more than 600 different writers [35]. The groundtruth information is provided at the word, line, and page levels allowing several types of experimentations. Overall, the database contains 115,320 labeled words that have been extracted using an automatic segmentation scheme and have been subsequently manually verified.

7.2 Competitions

Besides "static" databases, that are built with significant efforts and aim at a long term research, there are also several competitions that are run at conferences and workshops in the DAR field. The data collected by the organizers are usually available on the Internet. The format of the competitions is generally based on the distribution of training and test data to participants in the months before the conference. The data are freely available for research purposes, but a registration is usually required. The evaluation of the results for system comparison can be based on three main approaches. In some cases the participants send to the organizers the results obtained by their systems on a pre-defined test set. In other cases the participants send the executable programs to the organizers that compare the systems "in house". In the last approach the participants run their systems during the conference on new data previously unknown.

These competitions are helpful for the DAR research not only for the test data that are collected, but also for the development of performance evaluation methods as well as approaches for the automatic generation of synthetic data that are required to run these events. We list in the following some of the most recent competitions that have been organized at DAR conferences.

[7] http://marg.nlm.nih.gov/index2.asp
[8] http://www.iam.unibe.ch/∼ fki/iamDB

Page segmentation

Page segmentation is one important step in layout analysis and is particularly difficult when dealing with complex layouts. The page segmentation competition has been organized in the last ICDAR conferences (from 2001 onwards)[9]. The main objective of the competition is to compare the performance of layout analysis algorithms using digitized documents from common publications [36].

Arabic handwritten

Starting from the experience made with the IFN/ENIT-database a series of competitions to establish the state of the art of recognizing Arabic handwritten words has been organized at the ICDAR conference to give the opportunity to further develop methods and discuss results on Arabic recognition systems [37].

Symbol recognition

The GREC workshop is organized every two years by the IAPR Technical Committee on Graphics Recognition (TC 10). Graphics recognition is a subfield of DAR that deals with graphical entities in engineering drawings, maps, architectural plans, musical scores, mathematical notation, tables, and diagrams. In the last editions of the workshop some contests have been organized including an arc segmentation contest and a symbol recognition one[10]. The main goal of the latter contest is the comparison of various methods for recognizing linear graphic symbols, i.e. symbols made of lines, arcs and simple geometric primitives, which can be found in most graphic drawings.

Document image dewarping

In recent years the use of mobile devices for digitizing document images is becoming more and more important. One challenging task in the processing of camera-captured documents is the presence of page curl and perspective distortions. The goal of page dewarping is to flatten a document image in order to improve the recognition rate that can be achieved by state-of-the-art OCR systems. During the CBDAR 2007 workshop a page dewarping contest has been organized by the Image Understanding and Pattern Recognition group at the DFKI in Kaiserslautern (Germany). A database of camera captured documents with groundtruthed text-lines, text-zone, and ASCII text has been provided to participants[11].

[9] http://www.cse.salford.ac.uk/prima/ICDAR2007_competition
[10] http://symbcontestgrec05.loria.fr
[11] http://www.iupr.org/doku.php?id=didcontest

8 Concluding Remarks

In this introductory chapter we briefly introduced readers to the principal themes in the DAR research. We also attempted to present in a uniform way the main topics discussed in the chapters in the book that make, as a whole, an accurate picture of the current state of the art.

We hope that this book will contribute in two ways to the research in Document Analysis and Recognition. First, researchers in machine learning can contribute to the DAR research by starting from the problems addressed in this book and attempting to adopt innovative techniques previously studied in other domains. Second, we hope that researchers already active in the DAR field could employ machine learning techniques on several tasks modifying algorithms that are too frequently defined "by hand" by software developers.

References

1. Lyman, P., Varian, H.R.: How much information. Technical Report Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on 6-1-2007 (2003)
2. Sellen, A.J., Harper, R.: The myth of the paperless office. MIT press (2001)
3. Mori, S., Suen, C., Yamamoto, K.: Historical review of OCR research and development. Proc. IEEE **80** (1992) 1029–1058
4. O'Gorman, L., Kasturi, R.: Document Image Analysis. IEEE Computer Society Press, Los Alamitos, California (1995)
5. Casey, R.G., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Transaction on PAMI **18**(7) (1996) 690–706
6. Nagy, G.: Twenty years of document image analysis in PAMI. IEEE Transaction on PAMI **22**(1) (2000) 38–62
7. Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition: a comprehensive survey. IEEE Transaction on PAMI **22**(1) (2000) 63–84
8. Marinai, S., Gori, M., Soda, G.: Artificial neural networks for document analysis and recognition. IEEE Transactions on PAMI **27**(1) (2005) 23–35
9. Ha, T., Bunke, H.: Image processing methods for document image analysis. In: Handbook of character recognition and document image analysis. World Scientific (1997) 1–47
10. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks **2** (1989) 359–366
11. Taxt, T., Flynn, P.J., Jain, A.K.: Segmentation of document images. IEEE Transaction on PAMI **11**(12) (1989) 1322–1329
12. Nagy, G., Seth, S.: Hierarchical representation of optically scanned documents. In: Int'l Conference on Pattern Recognition. (1984) 347–349
13. Watanabe, T., Luo, Q., Sugie, N.: Structure recognition methods for various types of documents. MVA **6**(6) (1993) 163–176
14. Nagy, G., Seth, S., Viswanathan, M.: A prototype document image analysis system for technical journals. IEEE Computer **25**(7) (1992) 10–22
15. Kim, J.H., Kim, K.K., Suen, C.Y.: An HMM-MLP hybrid model for cursive script recognition. PAA **3**(4) (2000) 314–324

16. Gilloux, M., Lemarié, B., Leroux, M.: A hybrid radial basis function network/ hidden Markov model handwritten word recognition system. In: Int'l Conference on Document Analysis and Recognition. (1995) 394–397
17. Fu, L.M.: Neural networks in computer intelligence. McGraw-Hill, New York, NY (1994)
18. Jain, A.K., Zhong, Y.: Page segmentation using texture analysis. Pattern Recognition **29**(5) (1996) 743–770
19. Shih, F.Y., Chen, S.S.: Adaptive document block segmentation and classification. IEEE Trans. SMC **26**(5) (1996) 797–802
20. Strouthopoulos, C., Papamarkos, N.: Text identification for document image analysis using a neural network. Image and Vision Computing **16**(12/13) (1998) 879–896
21. Wang, J., Jean, J.: Segmentation of merged characters by neural networks and shortest path. Pattern Recognition **27**(5) (1994) 649–658
22. Lu, Z.K., Chi, Z., Siu, W.C.: Length estimation of digits strings using a neural network with structure based features. SPIE/IS&T Journal of Electronic Imaging **7**(1) (1998) 79–85
23. You, D., Kim, G.: An approach for locating segmentation points of handwritten digit strings using a neural network. In: Int'l Conference on Document Analysis and Recognition. (2003) 142–146
24. Pal, U., Sinha, S., Chaudhuri, B.: Multi-script line identification from indian documents. In: Int'l Conference on Document Analysis and Recognition. (2003) 880–884
25. Ishitani, Y.: Flexible and robust model matching based on association graph for form image understanding. Pattern Analysis and Applications **3**(2) (2000) 104–119
26. Dengel, A., Dubiel, F.: Clustering and classification of document structure -a machine learning approach-. In: Int'l Conference on Document Analysis and Recognition. (1995) 587–591
27. Cesarini, F., Lastri, M., Marinai, S., Soda, G.: Encoding of modified X-Y trees for document classification. In: Int'l Conference on Document Analysis and Recognition. (2001) 1131–1136
28. van Beusekom, J., Keysers, D., Shafait, F., Breuel, T.M.: Distance measures for layout-based document image retrieval. In: Proc. Second Int'l Workshop on Document Image Analysis for Libraries. (2006) 232–242
29. Kaski, S., Honkela, T., Lagus, K., Kohonen, T.: Websom self-organizing maps of document collections. Neurocomputing **21**(1–3) (1998) 101–118
30. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
31. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: Int'l Conference on Document Analysis and Recognition. (2003) 958–963
32. Suzuki, M., Uchida, S., Nomura, A.: A ground-truthed mathematical character and symbol image database. In: Int'l Conference on Document Analysis and Recognition. (2005) 675–679
33. Pechwitz, M., Maddouri, S.S., Magner, V., Ellouze, N., Amiri, H.: IFN/ENIT-database of handwritten arabic words. In: 7th Colloque International Francophone sur l'Ecrit et le Document. (2002)

34. Ford, G., Thoma, G.: Ground truth data for document image analysis. In: Symposium on Document Image Understanding and Technology. (2003) 199–205
35. Marti, U., Bunke, H.: A full english sentence database for off-line handwriting recognition. In: Int'l Conference on Document Analysis and Recognition. (1999) 705–708
36. Antonacopoulos, A., Gatos, B., Bridson, D.: ICDAR 2005 page segmentation competition. In: Int'l Conference on Document Analysis and Recognition. (2005) 75–79
37. Margner, V., Pechwitz, M., Abed, H.E.: ICDAR 2005 arabic handwriting recognition competition. In: Int'l Conference on Document Analysis and Recognition. (2005) 70–74

http://www.springer.com/978-3-540-76279-9

Machine Learning in Document Analysis and Recognition

Marinai, S.; Fujisawa, H. (Eds.)
2008, XII, 434 p. 142 illus., Hardcover
ISBN: 978-3-540-76279-9